AD_____

Award Number: MIPR 2MCTC32157

TITLE: The Toxin and Virulence Database: A Resource for Signature Development and Analysis of Virulence

PRINCIPAL INVESTIGATOR: Murray A. Wolinsky, Ph.D.

CONTRACTING ORGANIZATION: Los Alamos National Laboratory
Albuquerque, New Mexico 87185-5400

REPORT DATE: October 2004

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2004	Final (1 Sep 2002 - 1 Sep 2004)

4. TITLE AND SUBTITLE
The Toxin and Virulence Database: A Resource for Signature Development and Analysis of Virulence

5. FUNDING NUMBERS
MIPR 2MCTC32157

6. AUTHOR(S)
Murray A. Wolinsky, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Los Alamos National Laboratory
Albuquerque, New Mexico 87185-5400

E-Mail: murray@lanl.gov

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

In this joint effort with the University of Alabama at Birmingham, Walter Reed, MITRE and USAMRIID, we are developing a comprehensive database for microbial toxins and virulence factors (www.tvfac.lanl.gov). This database will contain all known toxins and virulence factors as well as their homologs. Each TVFac record and the associated gene records will be fully annotated to include all relevant sequence and structural information as well as biological, epidemiological, and clinical information. Human annotation, sequence-based data analysis, protein structure prediction, and natural language-based data mining will be integrated for a comprehensive understanding of the common mechanisms leading to pathogenesis. This effort is intended to provide necessary infrastructure for studying the common mechanisms of virulence and identifying common "themes" utilized by microorganisms to cause disease. Bioinformatic tools are being developed to allow users to explore the database and perform detailed analysis and to support detection and countermeasure development efforts. This effort is a continuation and expansion of the USAMRIID-funded effort for a "Cooperative Relational Database Initiative For Threat Reduction.

14. SUBJECT TERMS
No subject terms provided.

15. NUMBER OF PAGES
138

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover..... 1

SF 298..... 2

Introduction..... 4

Body..... 5

Key Research Accomplishments..... 13

Reportable Outcomes..... 15

Conclusions..... 16

References..... 17

Appendices..... 19

1. *Paper*: Murray Wolinsky, Jian Song, Jason Gans, Cathy Cleland, Jingao Dong, Robert Leach, Chris Stubben, Yan Xu, Kevin Anderson, Frank Labeda, Luther Lindler, Elliot Lefkowitz, Alexander Morgan, Marc Colosimo, Alexander Yeh and Lynette Hirschman, *TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004.
2. *Paper*: Jian Song, Yan Xu, Scott White, Kevin W. P. Miller and Murray Wolinsky. *SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004.
3. *Paper*: Yan Xu, Murray Wolinsky, Karla Atkins, and Jian Song, ***PhyloNeighborView**: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004.
4. *Paper*: Song,Jian, Carol Bonner, Murray Wolinsky, Roy A. Jensen. The TyrA family of aromatic-pathway hydrogenases in phylogenetic context. Submitted to *BMC Biology*.
5. *Paper*: Carol Rohl, Charlie E.M. Strauss, Dylan Chivian and David Baker. Modeling Structurally Variable Regions in Homologous Proteins with Rosetta, *Proteins: Structure, Function and Bioinformatics* **55**:656-677, 2004.
6. *Presentation*: Jason Gans. *Efficient Nucleic Acid Signature Development for Broad Spectrum Pathogen Detection, presented at TIGR Seventh Annual Conference on Computational Genomics*, Reston, VA, 21-24 October 2004.

Introduction

Sequence-related research in biothreat reduction has focused on two broad goals: signature development to support detection and forensics, and understanding virulence factors to support countermeasures and treatment. Although both areas of research are important, it is our sense that current efforts in the biothreat community are not balanced between these goals but favor signature development. This imbalance is due to a prevailing organism-specific orientation. We proposed to partially redress this imbalance by providing a resource that supports both signature development and analysis of the mechanisms of virulence in an evenhanded manner.

Here we report on our efforts to develop a resource (see http://www.tvfac.lanl.gov) that will ultimately include a comprehensive database of microbial toxins and virulence factors, together with associated analytical tools and products. This resource differs substantially from existing databases in its cross-pathogen orientation: the database and tools are organized around pathogenicity rather than organism.

While the contents of the proposed resource are novel, the structure is not. The configuration of capabilities is essentially identical to that presented in the (second-generation) annotated sequence databases that Los Alamos National Laboratory provides to the Department of Energy and to the National Institutes of Health.

Such a cross-pathogen resource will help improve existing signatures by helping relate these signatures to the pathogenic properties of organisms (making the signatures more robust as new "near-neighbors" are identified and also less easy to "engineer" away). It will also allow novel generic or "cross-pathogen" signatures to be developed.

Ideally, this resource will support the computational and experimental study of common themes and variations in microbial pathogenicity. By focusing on common properties, the resource will provide a foundation for uncovering the essential building blocks or "piece-parts" of pathogenicity. In so doing, it will support the development of more effective therapies and preventative measures. The cross-pathogen orientation of the resource also facilitates the development of novel tools as described below.

Quick and accurate diagnosis of the diseases caused by any biothreat agents or agents of public health concerns is critical for developing timely and effective responses against bioterrorist attacks. Our current efforts have been focused on detecting and responding to only a few well-known agents (e.g., *Bacillus anthracis*, *Yersinia pestis*, *Clostridium botulinum*, smallpox, *etc.*). This focus must be broadened to include other agents, including novel (engineered) ones.

Better understanding of toxins and virulence factors is not only important for development of counter measures but is also essential for accurate clinical diagnosis. A great number of toxins and virulence factors are structurally and functionally related and share conserved domains or signatures. Any recombinant toxin and/or virulence factor (henceforth abbreviated TVFac) or new TVFac produced by other less known pathogens will likely share some of the same conserved domains or signatures that have been found in a known TVFac. So understanding of the common features shared by all of the known TVFacs will help us in detecting and responding to new or engineered biothreat agents.

All of our current efforts have been species-oriented and focused on identifying species-specific signatures. We have also devoted considerable effort towards development of many specific genome sequence databases to facilitate these species-oriented efforts (http://www.cbnp.lanl.gov and http://www.stdgen.lanl.gov). However, there are few, if any, cross-species efforts made for the systemic study of microbial toxins. We believe a comprehensive microbial toxin database will provide the foundation needed for further computational and experimental analyses. Such database efforts will help us not only to develop methods for detecting a broad range of biothreat agents, but also to build counter-measures that may be effective against many biothreat agents that produce similar toxins and virulence factors.

Body

Our overall deliverable as specified by the statement of work is a resource that consists of:

- a comprehensive annotated microbial toxin and virulence factor database (data),
- a set of analytical tools providing novel analytical capabilities (software),
- studies summarizing results of analyses that involve the resource (publications), and
- a user-friendly interface to the resource (software/scripts).

Tasks and subtasks	*Deliverable type*	*UAB ID*	*Lead* LANL	UAB	Joint
1. Refine and expand comprehensive virulence factor database		1, 8			
a. Establish common schema	*Data*				
b. Populate records	*Data*				
c. Identify and establish relationships with appropriate curators	*Report*				
d. Integrate with related national and international efforts	*Report*				
2. Develop software tools for pathogenicity analysis and visualization		4, 6			
a. Develop HMMs for individual clusters	*Software*				
b. Predict pathogenicity islands	Software				
c. Develop improved methods of pathway data analysis	Software				
d. Develop visualization tools to aid in analyzing pathogenicity	Software				
3. Develop taxonomy of virulence factors		2, 5			
a. Cluster based on sequence similarity	*Data*				
b. Cluster based on inferred functional similarity	*Data*				
c. Cluster based on structural similarity	*Data*				
d. Analyze clusters	*Report*				
4. Derive cluster-based signatures	*Report*	4, 7			
5. Security		8			
Assess security requirements; ensure compliance with standards	*Report*				

Table 1. Tasks identified in proposed Statement of Work.

Overview

We have succeeded very well at some of these tasks, but others require further work. Here we present a brief overview of our accomplishments; a more detailed description follows immediately afterward.

The overall status of the primary task area, *1. Refine and expand comprehensive virulence factor database*, is largely on track. A resource, located at www.tvfac.lanl.gov, has been constructed and is publicly available. This resource has been re-engineered from our earlier work to address a number of identified shortcomings, Refer to the first paper in the Appendix, (*Wolinsky et al, TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004), for a more complete description of the contents and functionality of this resource. A new architecture was devised and implemented for reasons that are described in this work. The major shortcoming of the work is that the database is still inadequately populated. This is a much larger task than originally envisioned and requires substantial manual effort. However, efforts to secure follow-on funding through the Department of Homeland Security appear to be successful, and it is expected that this deficiency can and will be addressed. We have made substantial efforts to integrate with other efforts – the publication of the above report was one such effort, a joint book chapter was prepared and published (*Labeda et al*) and a new collaboration with researchers at the Defense Science

Technical Laboratory (Dstl) at Porton Down, UK has been inaugurated, largely to tailor the TVFac database to their needs. We are now applying this work to TSWG-funded efforts. In addition, a meeting has been scheduled in early November with Dr. Lefkowitz to completely integrate the work done at University of Alabama with that done at Los Alamos. This integration should be completed by the end of the calendar year.

The second task area, *2 Develop software tools for pathogenicity analysis and visualization*, has, on the whole greatly exceeded expectations. We have developed two "pipelines" for determining nucleic acid signatures, one based on identifying unique nucleic acid regions (and which leverages the Los Alamos developed mpiBLAST parallel BLAST software), and a second, polymorphism-based pipeline, that complements the above by finding SNPs and other polymorphisms for finer resolution signature candidate generation. While our results are still preliminary – and are undergoing experimental test now – the initial findings are very encouraging. A key goal of this effort is to uncover *generic* signatures for pathogens. Our pipeline was applied to a set of diverse enteric pathogens of commercial interest to a collaborator. Over 125 potential signature candidates were identified – as being common to this diverse set and absent from other organisms – and an analysis of these candidates suggest that many may be connected with specific regulatory apparati. This is an encouraging result, because it suggests that (a) generic signatures exist; (b) these signatures have biological significance; and (c) these signatures may be robust and not easily "evolved" or engineered away. If these results are confirmed, they will represent important progress for biothreat detection, forensic analysis, and ideally countermeasure development. An easy-to-use polymorphism-based signature pipeline was also constructed to complement the unique regions work. This pipeline is focused on identifying variability for strain resolution as opposed to commonality. It is based on comparative analysis of neighboring genomes and has been applied to identifying SNPs in many pathogenic organisms. The most interesting results – also preliminary – have been obtained by a collaborator at Los Alamos (Dr. Scott White) who has used this tool to identify seven new SNPs for *Bacillus anthracis*, one of the most monomorphic pathogens known. These results are undergoing laboratory testing and will be published shortly. We have developed novel and user-friendly applications for visualization and pathogenicity island prediction. The results are described in detail in (Wolinsky et al) and are expanded on below.

The third task area, *3. Develop taxonomy of virulence factors*, is incomplete. Dr. Jian Song constructed and reported on a manually-generated taxonomy of these factors (*Wolinsky et al*). Nucleic-acid cluster were generated by Dr. Gans and reported on at an earlier status meeting at USAMRIID. However, protein-based and literature derived clusters have not yet been generated. In the case of literature-derived clusters, we have entered a collaboration with Dr. Lynette Hirschman at MITRE, to help us derive these clusters and this work remains in progress. For the protein-structure based clusters, we have developed the tools and infrastructure to perform this work, but have only begun to apply them. In large part, our previously existing software, Rosetta, required too much manual intervention and our existing hardware (a 240 node Linux cluster) has not been reliable enough nor fast enough to perform the requisite genome wide predictions of structure required. We have worked on improving the codes (*Rohl et al*, included), but more innovative approaches are necessary. The funding derived from this work assisted in the complete automation of the Rosetta protein structure algorithm, (the fully automated pipeline is called Robetta), which is available on our cluster and publicly at www.robetta.org. In addition, agreements with IBM have been made to make Robetta available as a screen-saver on PC desktop machines to vastly increase the potential computational power applicable. This screensave should be available this November (2004). We intend to complete the genome wide protein structure prediction for all NIH Category A and B pathogens and make these results available through our site. Further work is required to accomplish our goals in this area and this work is continuing using other funds.

The fourth task area, *4. Derive cluster-based signatures*, has been addressed in ways we believe exceed the original plans, but which took an alternate approach. Specifically, the pipelines described above have produced signature candidates that may meet all expectations regarding usability and biological significance (see above). These generic signatures have been produced for a diverse set of enteric pathogens and are now being produced for all NIH Category A and B pathogens. Further computational and experimental investigations are in progress.

The fifth and final task area, *5. Security*, has largely been deferred. We have ensured that our database and web site conforms to all Department of Defense guidance and to both Department of Energy and Los Alamos National Laboratory requirements. However, there is currently no sensitive data or capabilities on the site and no current need for specific security measures.

Discussion

Our most exciting accomplishment has been the

1. Construction and operation of a novel unique regions nucleic signature development pipeline

This pipeline is nucleic-acid based and identifies potential signature elements for groups of particular pathogenic organisms. This pipeline was built on top of the Los Alamos-developed mpiBLAST implementation and runs of a 240 node Linux cluster located at Los Alamos National Laboratory. mpiBLAST is freely available open source software and can be obtained at http://mpiblast.lanl.gov. Unlike other approaches to signature development, the Los Alamos pipeline does not require alignment of the target genomes. This freedom from imposed alignments makes the pipeline more flexible than competing approaches and is particularly appropriate when looking for *generic* signatures, as opposed to species-based signatures. The construction of this pipeline goes a long way to satisfying one of the key deliverables of the proposal.

The pipeline work has been presented by Dr. Jason Gans at the TIGR Seventh Annual Conference on Computational Genomics in Reston, VA on 21-24 October 2004. The slides Dr. Gans presented are included in this document. A more complete paper is in preparation. Currently the pipeline is being exercised on all NIH Category A and B pathogens. The signatures which emerge from this work will be made available through our TVFac website. Preliminary work, which was reported at the TIGR meeting, was performed on the set of commercially-interesting enteric pathogens shown in Table 2 below.

E. coli 042
E. coli CFT073
E. coli E2348 69
E. coli O157H7
E. coli K12
E. coli O157H7 EDL933
S. typhi
S. bongori 12419
S. enteritidis PT4
S. typhimurium LT2
S. typhi Ty2
S. gallinarum 287 91
S. typhimurium DT104
S. flexneri 2a
S. dysenteriae M131649
S. flexneri 2a 2457T
S. typhimurium SL1344
S. sonnei 53G
Y. pestis biovar Mediaevails
Y. pestis CO92
Y. pestis KIM
Y. pseudotuberculosis
Y. enterocolitica 8081

Table 2. Enteric pathogens used for proof-of-principle generic pathogen signature candidate development

Running the pipeline on this set of pathogens produced 125 signature candidates (unique nucleic acid regions) that were common to all of these enteric targets. Primers have been synthesized and experimental verification of these predictions is in progress. The preliminary results appear promising. Scientifically, the most interesting result – which is still preliminary – lies in the distribution of these signatures. Table 3 presents the distribution.

	Inside gene	Outside gene (intergenic)	Spanning intergenic space/gene boundary
Expected	87%	10%	3%
Observed	55%	18%	26%

Table 3. Distribution of 125 enteric pathogen signature candidates. Signature candidates were classified as being entirely inside a gene, entirely outside (intergenic), or spanning a gene boundary. The expected numbers assume a uniform distribution of candidate placement. Note that almost nine times as many candidates span boundaries as would be expected by chance.

The most suggestive finding is that disproportionately many (approximately nine times chance) signature candidates cross genomic/intergenic boundaries. This immediately suggests these candidates may be involved in gene regulation. This hypothesis is strengthened by additional evidence. First, the signatures are preferentially located near a gene start. (See figure 1.)

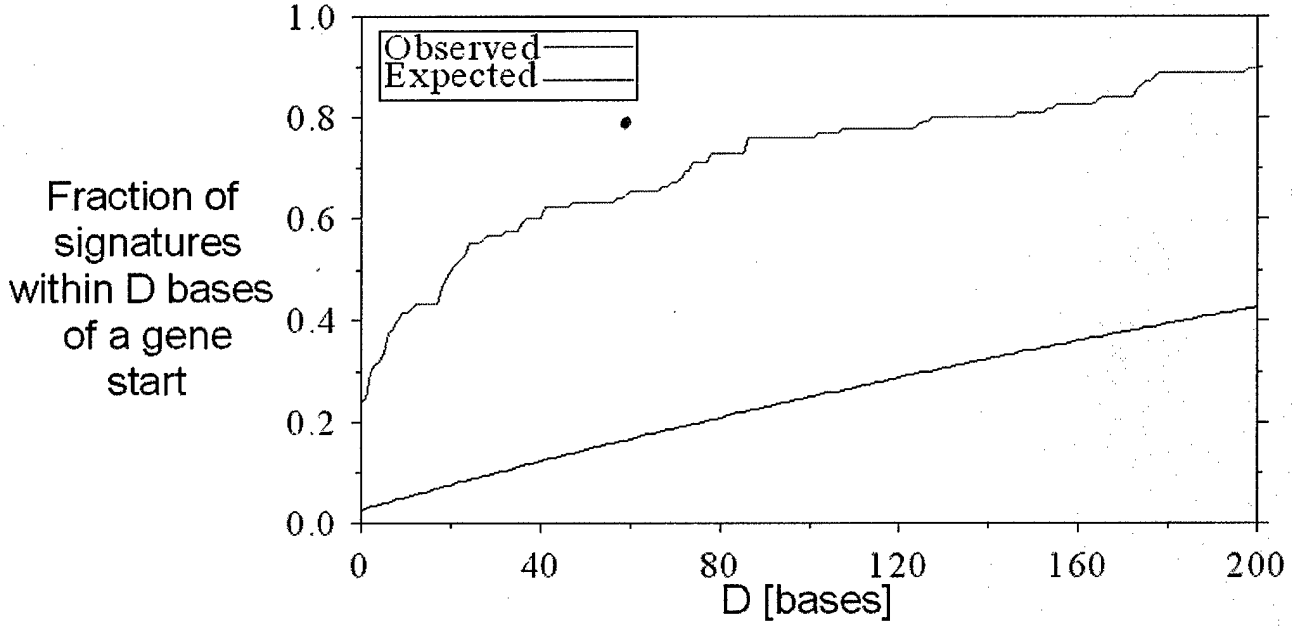


Figure 1. Signature candidates are preferentially located near gene starts.

Second, informatic analyses show a correlation between the candidate signatures and known regulatory elements, including transcription factor binding sites and ribosomal binding sites. Therefore our preliminary analyses suggests that generic signatures are likely to be associated with regulatory apparatus. These results give hope that identified signatures will not be easy to engineer away.

More detailed computational and experimental investigations are underway to determine how robust these results are. We will publish these results more formally after more detailed computational and experimental data are available.

The unique regions pipeline described above is useful for broad detection of pathogens or other microbes. It is also useful to have diagnostic (or forensic) signatures that are capable of resolving fine strain differences. To that end we have also developed a tool, a web site (www.snpsfinder.lanl.gov), and published our results (see Appendix, paper 2) for

2. A polymorphism-based nucleic signature development pipeline

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variations among closely related microbial species, strains or isolates. Some SNPs confer selective advantages for microbial pathogens during infection and many others are powerful genetic markers for distinguishing closely related strains or isolates that could not be distinguished otherwise. To facilitate SNPs discovery in microbial genomes, we have developed a web-based application, *SNPsFinder*, for genome-wide identification of SNPs. *SNPsFinder* takes multiple genome sequences as input to identify SNPs within homologous regions. It can also take contig sequences and sequence quality scores from on-going sequencing projects for SNPs prediction. *SNPsFinder* will use genome sequence annotation if available and map the predicted SNPs regions to known genes or regions to assist further evaluation of the predicted SNPs for their functional significance. *SNPsFinder* can generate PCR primers for all predicted SNPs regions according to user's input parameters to facilitate experimental validation. The results from *SNPsFinder* analysis will be accessible through the World Wide Web. The *SNPsFinder* program is available at http://snpsfinder.lanl.gov/.

This tool has been applied by Dr. Scott White to generate new candidate polymorphisms in *Bacillus anthracis*, which is a notoriously monomorphic organism. Current results have identified 7 to 13 new polymorphisms which are being subjected to laboratory verification. The tool is described in the attached paper, Song *et al. SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004. Funding from the FBI assisted us in performing this work.

3. Establishment of a prototype database/website

This site has been established at Los Alamos at www.tvfac.lanl.gov. A recent publication describing this work in included below (*Wolinsky et al*). This publication describes the subgoals of

3a. Development of a functional classification of pathogens

3b. Development of an architecture for a novel cross-pathogen resource

Additional work, not reported in that publication is presented in the attached manuscript, submitted to *BMC Biology* on an informatics-based analysis of the TyrA family of aromatic-pathway dehydrogenases. (Song,Jian, Carol Bonner, Murray Wolinsky, Roy A. Jensen. The TyrA family of aromatic-pathway hydrogenases in phylogenetic context. Submitted to *BMC Biology*.) This work is intended as an initial example of the types of data and analysis that the TVFac will eventually support.

4. Application of the TVFac database and tools to a challenging bioinformatics problem

The metabolic pathways of an organism represent its metabolic capability and determine its range of environment or hosts where it can grow, survive and cause infections. Identification of unique pathways through comparative genomic analysis will help us to better understand the difference in virulence and pathogenesis between different bacterial species. One such example is recently discovery that the presence of trytophan recycling pathway in non-pathogenic *Chlamydophila psittaci* when compared with the pathogenic chlamydiae (*Chlamydia trachomatis, C. muridarum* and *Chlamydophila pneumoniae*) (Xie, et al., Genome Biology 2002, **3**:research0051.1-0051.17). Tryptophan limitation caused by production of interferon-? by the host and subsequent induction of indoleamine dioxygenase is a key aspect of the host-parasite interaction. It is because of the lack of this recycling pathway that the pathogenic chlamydiae have learned to recognize tryptophan depletion as a signal that leads transition to the persistent (chronic) state of pathogenesis. There are many examples that presence or absence of unique pathways leads to the development of unique natural countermeasures. Therefore comparative pathway analysis will provide new insights and help us in our efforts to develop more effective countermeasures.

The availability of genome sequences has provided us new opportunities for comparative pathway analysis. We are not longer limited by the availability of sequence data, but rather by the interpretation of those data. Errors in current annotations are abundant. Since new annotations rely upon the existing set of annotations, errors have proliferated and continue to proliferate. The situation is sufficiently severe that in many cases only experts familiar with particular pathways can recognize and sort out errors. A major problem has been the untidy and erratic nomenclature used to identify genes, e.g., (i) use of the same name for different genes, (ii)

use of the same name for genes that, on the one hand, encode single-domain proteins (due to gene fusion), and, on the other hand, genes that encode multi-domain proteins, and (iii) use of different names in different organisms for genes encoding the same protein. It is absolutely critical (and surely inevitable) that a logical and consistent universal nomenclature be established.

We started the AroPath to complement the TVFAac database. AroPath is taking the initiative that this situation can best be addressed with a comprehensive, expert-assisted manual effort that is undertaken with a realistically manageable subsystem of metabolism. In our case this subsystem is the extensive network of aromatic metabolism. No matter how high the quality of an analysis is at a given time, it will become outdated rapidly as new genomes come on line. Therefore, our approach is to produce tools which are able to lock in the current advances in analysis as they come, which are freely available and interactive so that the previous work can be efficiently exploited, and which is amenable to progressive updating and refinement via curator approval at AroPath. We hope the AroPath will be further developed and become a model system for other metabolic pathways, eventually to entire metabolic pathways to facilitate system biology studies and modeling of single organism or a community of organisms.

5. Development of novel tools for visualization

Discussions at USAMRIID revealed the need for a new type of visualization tool – one that would allow viewing multiple genomes simultaneously, preserve phylogenetic relationships, and allow gene neighborhood organization to be displayed. We constructed such a tool and have reported on it in the attached paper, Yan Xu *et al PhyloNeighborView: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004. This paper describes the development of PhyloNeighborView, a client-side, cross-platform pure Java visualization tool that integrates 16S-based phylogenetic tree with annotated genomes. We have integrated the completely sequenced genomes from TIGR CMR into PhyloNeighborView. CMR was chosen because of its robust genome annotation. PhyloNeighborView allows visualization of gene-level comparisons across multiple species and uses the phylogenetic tree as a guide to display the gene neighborhood and analyze the level of gene conservation across tree nodes. Advantages of PhyloNeighborView over other software are (1) integration of phylogenetic tree to allow comparative genomics in phylogenetic context, (2) addition of sequence similarity-based clustering to allow users to browse multiple homologs (paralogs), (3) genes are color-coded according to their predicted biological functions, which will help to understand the functions of uncharacterized genes in the gene neighborhood, (4) biologist-friendly, no data manipulations are required and users simply download and run PhyloNeighborView locally like any other desktop application.

6. Development of a wavelet-based tool for candidate pathogenicity island prediction

The identification of possible pathogenicity islands is one of the goals of this effort. Understanding lateral gene transfer is important in understanding and analyzing potential signatures of pathogenicity abd in uncovering evidence of genetic engineering. This goal has led us to produce the genWave tool, developed by Robert Leach at Los Alamos, and available through the TVFac website. The tool is described in Wolinsky *et al*. Recent work on *Yersinia pestis* signatures (in progress, by Chris Stubben) has highlighted the importance of this research. In particular, a pathogenicity island analysis performed by Mr. Stubben of signatures recently reported on by Patrick Chain *et al* in *Proc. Nat. Acad. Sciences*, has shown that presence of a signature element on an island is a good predictor that such a signature will *not* be robust. This research uncovered one exception to this rule, which is also of interest. A single island, associated with pathogenicity, appears to be a good source of signature candidates for *pathogenic Y. pestis as opposed to Y. pestis per se.* This pathogenicity island lacks elements that would assist in mobility of the element, which is potentially why the signature remains robust. Summarizing, this analysis has shown that signature candidates on mobile islands are unlikely to be robust, whereas candidates on islands that have lost mobility may be good predictors of pathogenicity. More detailed analysis is in progress, with results expected to be delivered to USAMRIID in early November 2004.

Some of the analysis that has gone into these results is presented below. This discussion also suggests future research that may be of benefit. *Yersinia pestis* has a number of potential pathogenicity islands that have been identified by various means. The most reliable methods are annotation-based, such as the identification of virulence determinants (that are implied to have come from horizontal transfer), the identification of highly

conserved tRNA sequences (which have been found to commonly flank regions of horizontal transfer), the presence of IS elements or integrases (which facilitate horizontal transfer), and the presence of phage proteins. There is only one purely sequence-based computational method which has been regularly applied to suggest the presence of horizontal transfer regions: GC content. None of any of these methods by themselves are enough to qualitatively identify a horizontal transfer, hence the more evidence one can gather to aid in identification, the better.

We have performed a pathogenicity island analysis on *Y. pestis* and have uncovered further support in the identification of one of its potential pathogenicity islands ("Genome sequence of *Yersinia pestis,* the causative agent of plague" *Nature*, ***413***, 523) using raw chromosomal sequence analysis. The most common sequence analysis method used to suggest the presence of a horizontal transfer is abnormal fluctuations in GC content. The premise is that different genomes have different inherent GC contents.

DNA composition can be influenced by a number of factors: environment (such as access to GC bases), the way replication (or other) machinery works such as is believed to be the case in GC skew (which is very profound in *Y. pestis*), evolutionary pressure, etc. The characteristics we will focus on here are structural, namely the stiffness of DNA. It is possible for example, that a cell's molecular machinery (e.g. for replication) could work more efficiently if the DNA being replicated had a particular shape or flexibility. Therefore, some mutations could be favored over others, by either easier access of certain bases in certain situations or easier manipulation of them for proper orientation. Given that different trinucleotides have different degrees of stiffness/flexibility that have been experimentally measured (*Gromiha*), it is reasonable to imagine that (as in the case of GC content) structural characteristics could influence the composition of DNA in some organisms. DNA from different organisms will have different characteristics based on the varying influences.

One inherent impediment in this analysis of course is the age of the foreign DNA segments. The older a foreign piece of DNA is, the more it is subject to the native DNA composition influences and the more the foreign DNA takes on the characteristics of the surrounding DNA. So the more information we can extract from the remaining foreign sequences, the better our ability to identify them. There is one technique we can employ to reduce native 'noise' gradually introduced by the host organism's composition influences. The application of the Daubechies wavelet transform emphasizes significant foreign DNA peaks and valleys by eliminating the high frequency fluctuations (noise) and isolating the low frequency shifts in DNA stiffness or GC frequencies.

Results presented here show that using DNA stiffness as a distinguishing organism-specific sequence characteristic captures all but one of the significant peaks and valleys (as determined by whether or not a large peak/valley (greater than 3 standard deviations) corresponds with a potential pathogenicity island identified by annotation methods) shown in a GC content analysis and additionally captures another pathogenicity island that GC content analysis misses completely. Stiffness analysis also shows two other moderately sized (2 standard deviations) events that correspond with 2 additional potential pathogenicity islands that GC misses.

Of the 21 potential pathogenicity islands, wavelet analysis with genWave shows that GC analysis has 7 peaks/valleys (greater than 2 standard deviations) that occur in or directly adjacent to potential pathogenicity islands while the DNA Stiffness measure has 9.

The correlation between DNA stiffness and GC content seen in the graphical analysis is strong despite the fact that very little correlation can be seen between the measurements for each individual trinucleotide and their GC content. The biggest significant difference between the two graphs is the valley at around 4.1MB where there is an annotated potential pathogenicity island containing insecticidal toxin complex genes with hits to *E. coli* and *S. typhimurium*, a putative exported protein, two phage related proteins, and a pair of transcriptional regulatory proteins at one end that also have hits to *E. coli* and *S. typhimurium*. All the reading frames are in the same orientation. GC analysis shows only a modest valley at this position that does not stand out among neighboring peaks and valleys. However the DNA stiffness valley is much more pronounced. In the genWave analysis, it deviates by over 3 whole standard deviations whereas the number of deviations in the GC analysis is 10 times less.

Another moderate, yet notable difference corresponding with a potential pathogenicity island occurs at 1.24MB. Here we see a peak in the stiffness analysis that is absent in the GC analysis. This region is annotated to be a phage remnant with a neighboring conserved tRNA-Asp.(1)

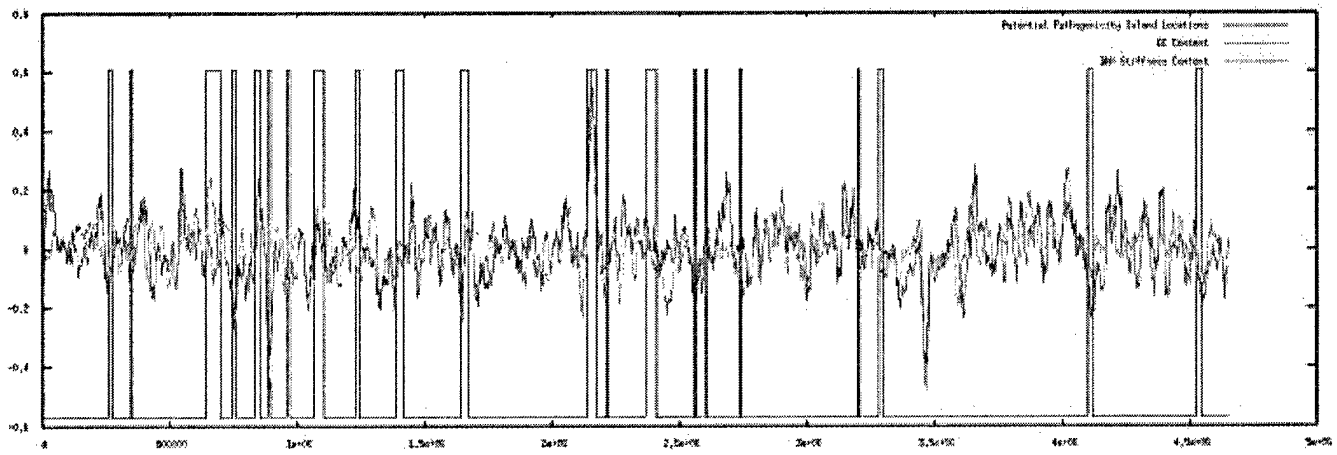


Figure 2.GC and DNA Stiffness with potential pathogenicity islands locations marked for comparison. Both GC and DNA Stiffness were fit into a range of 0 to 1 and then normalized. A 20KB window was slid at increments of 2KB (accounting for the circularity of the genome) and each point represents the center of the window.

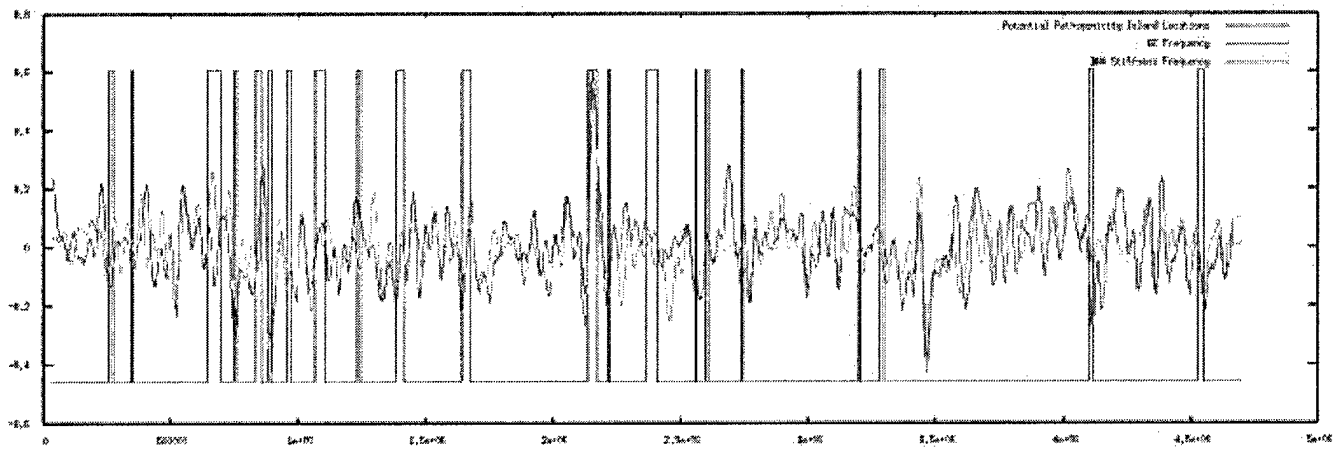


Figure 3. GC and DNA Stiffness with potential pathogenicity islands locations marked for comparison. Both GC and DNA Stiffness were fit into a range of 0 to 1 and then normalized. Using the genWave(3) tool, a signal was generated using a 10B window containing the average number of GCs and DNA Stiffness coefficients. The signal was then padded and transformed with the daubechies wavelet. The low frequency component was isolated and reverse transformed.

These results are particularly telling as to the factors involved in inherent DNA compositional influences and could reveal important mechanisms behind the direction of evolution in the microbial world.
The key question is why there is a link between DNA stiffness/flexibility and the base composition of DNA. Not only could this be due to the workings of the replication machinery, but the influence could also be downstream of replication, such as DNA repair mechanisms or advantageous DNA binding protein efficiencies. DNA stiffness/flexibility is also only one structural aspect that may be involved. For example, structural characteristics like degrees and energies of twist, turn, tilt, roll, shift, slide, rise, or other combinations of characteristics could be important, not only in compositional influences, but in specific binding or other mechanisms. Since we know that there are strong compositional influences in *Y. pestis* as is evident by its profound GC skew, using this organism to perform graphical analysis based on other structural features is the next step in deciphering what could be causing the observed compositional influences.

DNA stiffness is definitely not all of the picture. There are some places, just as in GC analysis, where statistically significant peaks and valleys based on DNA stiffness do not correspond to any annotated potential pathogenicity islands or for that matter, any horizontal transfers. One explanation could be that a combination of composition influences is what denotes significance and identifies an island. Another reason could be that stiffness or flexibility in a particular region serves a functional purpose. We are continuing to attempt to understand how best to predict pathogenicity islands to utilize these predictions.

Key Research Accomplishments

- ***Construction and operation of a novel unique regions nucleic signature development pipeline***

The pipeline work has been presented by Dr. Jason Gans at the TIGR Seventh Annual Conference on Computational Genomics in Reston, VA on 21-24 October 2004. This work was jointly supported by the DHS. This initial report does credit DoD funding: however, a final publication will fully acknowledge the support of the DoD in this work.

- ***A polymorphism-based nucleic signature development pipeline***

The *SNPsFinder* program is available at http://snpsfinder.lanl.gov and was published in Song *et al. SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004. This work was jointly supported by the FBI and the DHS. An accepted publication acknowledges the support of this DoD effort, and the dedicated website also credits this effort.

- ***Development of generic signatures of pathogenicity***

The accomplishments are summarized in the Gans presentation. Again, the funding received through this effort was not fully acknowledged, which we shall correct in a future publication.

- ***Establishment of a prototype database/website***

This site has been established at Los Alamos at www.tvfac.lanl.gov. A recent publication describing this work in included below (Murray Wolinsky, Jian Song, Jason Gans, Cathy Cleland, Jingao Dong, Robert Leach, Chris Stubben, Yan Xu, Kevin Anderson, Frank Labeda, Luther Lindler, Elliot Lefkowitz, Alexander Morgan, Marc Colosimo, Alexander Yeh and Lynette Hirschman, *TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004). This publication also describes the subgoals of

  - ***Development of a functional classification of pathogens***
  - ***Development of an architecture for a novel cross-pathogen resource***

The website was developed almost completely with DoD support as the key deliverable of this effort. Both the website and the publication acknowledge the DoD role in establishing the resource. Future development of the site will continue through the DHS if we are unable to attract additional DoD support. Both the website and the publication above acknowledge DoD support.

A pilot study on the TyrA family of dehydrogenases was conducted using the resources provide by the DoD. This study has been submitted to BMC Biology and acknowledge the contributions of the DoD. A website was also set up for this effort, http://snp.lanl.gov/AroPath/SupplMaterials/TyrA/Table6.html, which is linked to the TVFac site and acknowledges the DoD contributions.

- ***Development of novel tools for visualization***

This tool is available at www.tvfac.lanl.gov and is presented in Yan Xu *et al PhyloNeighborView: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004. This work has been submitted for publication and acknowledges the funding provided by the DoD to this effort. This work was also co-funded by the DHS.

- ***Development of a wavelet-based tool for candidate pathogenicity island prediction***

A user-friendly application, *genWave*, a pathogenicity island finder developed by Robert W. Leach is available at www.tvfac.lanl.gov and is described in Wolinsky *et al.* The support provided by this effort is acknowledged in the website and will also be acknowledged in a future publication, in preparation.

- ***Full automation of the Robetta protein structure pipeline***

This tool is available through our cluster and at www.robetta.org. The Robetta effort has derived funding through a number of efforts. We will ensure that the DoD support for this effort will be acknowledged on the Robetta server.

Reportable Outcomes

The principal outcome of this project has been to establish a new architecture for the database and a usable, though highly incomplete, web site at www.tvfac.lanl.gov. This website provides access to the tools described earlier.

Several publications have resulted from this work. The following publications are included within this report:

1. *Paper*: Murray Wolinsky, Jian Song, Jason Gans, Cathy Cleland, Jingao Dong, Robert Leach, Chris Stubben, Yan Xu, Kevin Anderson, Frank Labeda, Luther Lindler, Elliot Lefkowitz, Alexander Morgan, Marc Colosimo, Alexander Yeh and Lynette Hirschman, *TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004.
2. *Paper*: Jian Song, Yan Xu, Scott White, Kevin W. P. Miller and Murray Wolinsky. *SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004.
3. *Paper*: Yan Xu, Murray Wolinsky, Karla Atkins, and Jian Song, ***PhyloNeighborView**: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004.
4. *Paper*: Song,Jian, Carol Bonner, Murray Wolinsky, Roy A. Jensen. The TyrA family of aromatic-pathway hydrogenases in phylogenetic context. Submitted to *BMC Biology*.
5. *Paper*: Carol Rohl, Charlie E.M. Strauss, Dylan Chivian and David Baker. Modeling Structurally Variable Regions in Homologous Proteins with Rosetta, *Proteins: Structure, Funciton and Bioinformatics* **55**:656-677, 2004.

A book chapter (not included) was also prepared as part of this effort.

6. Lebeda, F.J., Wolinsky, M. and Lefkowitz, E.J. "Information Resource and Database Development." In *Biological Weapons Defense: Principles and Mechanisms of Infectious Disease*, Lindler, L., Lebeda, F. J. and Korch, G., eds., Humana Press, Inc., Totowa, New Jersey (in press 2004).

In addition, the following preliminary oral presentation (also included) describing this work was made:

7. *Presentation*: Jason Gans. *Efficient Nucleic Acid Signature Development for Broad Spectrum Pathogen Detection, presented at TIGR Seventh Annual Conference on Computational Genomics*, Reston, VA, 21-24 October 2004.

This effort partially funded the post-doctoral work by Dr. Jason Gans, who is expected to be offered a full-time position at Los Alamos National Laboratory, in part due to this effort.

Funding from the FBI was secured to assist in the development of the SNPsFinder application, part of our signature pipeline capability.

Additional funding from TSWG has been obtained to continue the development of our pipelines and to apply them to all NIH Category A and B pathogens.

A collaboration with IBM was entered into to develop a screensaver version of Robetta to make high-quality protein structure prediction freely and widely available. This screensaver should be available in November 2004.

Dr. Lynette Hirschman, at the MITRE Corporation, was awarded a substantial internal research grant over a period of three years in large part to work with us to improve this resource.

Follow-on funding to continue this effort has been secured from the Department of Homeland Security as part of a collaborative US/UK effort with Porton Down researchers to investigate generic signatures and countermeasures. Additional funding is expected.

Conclusions

We believe we have made a good start at establishing the proposed resource. TVFac is available to the public at http://www.TVFac.lanl.gov and most of the tools developed for the database are also available to allow users to perform real-time and online analysis of their own genomic sequence data. The current version is mostly proof-of-concept and preliminary: much work is still required to make TVFac a useful resource. ***Above all, we need to devote manual effort and develop more clever automated ways to populate as quickly, accurately and completely as possible the database.***

Our current effort built a good infrastructure to allow expansion and community participation. ***We will need to add new types of data, including expression data and proteomic data as this data becomes available.*** A start has been taken in this regard by developing the collaboration with researchers at Dstl, Porton Down. We must further develop new comparative capabilities to allow us to predict virulence mechanisms and pathogenesis of new and emerging pathogens. Above all, we want to encourage interested users to participate in developing an invaluable resource for the entire biodefense community.

T***he computational proteomics work should be advanced further.*** We intend to add experimental or predicted structure information on every gene. Some of this work will emerge naturally from the foundations we have established. Additional support for this effort may be required, however.

The development of generic signatures appears to be very promising: further work should be pursued. Partial support for such work has been obtained from Department of Homeland Security, from the United States Department of Agriculture, and from TSWG. But much work needs to be done. It may be the case, that for the first time, we can start to ***understand the science underlying signature development***. Our current work suggests that understanding regulatory apparatus and lateral gene transfer will play a role in this science. It looks like good signatures, particularly good generic signatures, may be regulatory elements: ***we need to be able to computationally predict and analyze these regulatory elements***. It also looks like robust signatures will avoid pathogenicity islands, unless these islands have been immobilized. ***We need better tools for analyzing lateral gene transfer events***.

References

Andrade, M.A., *et al.*, Automated genome sequence analysis and annotation. *Bioinformatics*, 1999. 15(5): p. 391-412.

Drell, S.D., A.D. Sofaer, and G.D. Wilson, *The new terror: facing the threat of biological and chemical weapons*. Stanford, CA. Hoover Institution Press, 1999.

Feng, W.-c., The Design, Implementation, and Evaluation of mpiBLAST. *ClusterWorld*, 2003.Finlay, B.B. and S. Falkow, Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev*, 1997. 61(2): 136-69.

Fitch, J.P., et al., Biosignatures of pathogen and host. *Proc IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, NC 2002.

Forslund, D. Experiences Developing Distributed Object Applications. *OOPSLA* 1998. 1988. Vancouver, BC.

Gans, Jason. "Efficient Nucleic Acid Signature Development for Broad Spectrum Pathogen Detection", presented at TIGR Seventh Annual Conference on Computational Genomics, Reston, VA, 21-24 October 2004.

Gromiha, M. Michael, "Structure Based Sequence Dependent Stiffness Scale for Trinucleotides: A Direct Method", Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Hirschman, L., Morgan, A. and Yeh, A., Rutabaga by Any Other Name: Extracting Biological Names. *J. Biomed. Inform.*, 2002. 35(4): p. 247-259.

Kanehisa, M.I., Los Alamos sequence analysis package for nucleic acids and proteins. *Nucleic Acids Res*, 1982. 10(1): p. 183-96.

Karlin, S., Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol*, 2001. 9(7): p. 335-43.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya, Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 2000. 288(5472): p. 1789-1796.

Lebeda, F.J., Wolinsky, M. and Lefkowitz, E.J. "Information Resource and Database Development." In *Biological Weapons Defense: Principles and Mechanisms of Infectious Disease*, Lindler, L., Lebeda, F. J.and Korch, G., eds., Humana Press, Inc., Totowa, New Jersey (in press 2004).

Morgan, A., Hirschman, L., Yeh, A. and Colosimo, M. Gene Name Extraction Using FlyBase Resources. *ACL Workshop on Natural Language Processing in Biomedicine*. 2003. Sapporo, Japan.

Pietro *et al.* Finding Pathogenicity Islands and Gene Transfer Events in Genome Data, *Bioinformatics*, 2000.

Riley, M., Functions of the gene products of *Escherichia coli. Microbiol Rev*, 1993. 57(4): p. 862-952.

Carol Rohl, Charlie E.M. Strauss, Dylan Chivian and David Baker. "Modeling Structurally Variable Regions in Homologous Proteins with Rosetta", *Proteins: Structure, Function and Bioinformatics* **55**:656-677, 2004.

Song,Jian, Carol Bonner, Murray Wolinsky, Roy A. Jensen. "The TyrA family of aromatic-pathway hydrogenases in phylogenetic context." Submitted to *BMC Biology*.

Song, Jian, Yan Xu, Scott White, Kevin W. P. Miller and Murray Wolinsky. *SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004.

Wolinsky, Murray, Jian Song, Jason Gans, Cathy Cleland, Jingao Dong, Robert Leach, Chris Stubben, Yan Xu, Kevin Anderson, Frank Labeda, Luther Lindler, Elliot Lefkowitz, Alexander Morgan, Marc Colosimo, Alexander Yeh and Lynette Hirschman, *TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004.

Xie, G., *et al*., Evolutionary origin of operonic genomic segments containing redundant copies of tryptophan-pathway genes: one in *Xyllella* species and the other in heterocystous cyanobacteria. *Genome Biol*, 2002. 4(R14).

Xu, Yan, Murray Wolinsky, Karla Atkins, and Jian Song, **PhyloNeighborView***: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004.

Yeh, A., L. Hirschman, A. Morgan, Evaluating text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, 2003. 19(Supplement 1): p. 331-339.

Appendices

1. *Paper*: Murray Wolinsky, Jian Song, Jason Gans, Cathy Cleland, Jingao Dong, Robert Leach, Chris Stubben, Yan Xu, Kevin Anderson, Frank Labeda, Luther Lindler, Elliot Lefkowitz, Alexander Morgan, Marc Colosimo, Alexander Yeh and Lynette Hirschman, *TVFacDB – A Comprehensive Microbial Toxin and Virulence Factor Database*, Proceedings of the BTR 2004 Unified Science and Technology for Reducing Biological Threats and Countering Terrorism, Albuquerque, NM, 17-19 March 2004.
2. *Paper*: Jian Song, Yan Xu, Scott White, Kevin W. P. Miller and Murray Wolinsky. *SNPSFinder – A Web-Based Application for Genome-Wide Discovery of Single Nucleotide Polymorphisms in Microbial Genomes*, submitted to BMC Bioinformatics, October, 2004.
3. *Paper*: Yan Xu, Murray Wolinsky, Karla Atkins, and Jian Song, ***PhyloNeighborView**: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms*, submitted to *Bioinformatics*, October, 2004.
4. *Paper*: Song,Jian, Carol Bonner, Murray Wolinsky, Roy A. Jensen. The TyrA family of aromatic-pathway hydrogenases in phylogenetic context. Submitted to *BMC Biology*.
5. *Paper*: Carol Rohl, Charlie E.M. Strauss, Dylan Chivian and David Baker. Modeling Structurally Variable Regions in Homologous Proteins with Rosetta, *Proteins: Structure, Function and Bioinformatics* **55**:656-677, 2004.
6. *Presentation*: Jason Gans. *Efficient Nucleic Acid Signature Development for Broad Spectrum Pathogen Detection, presented at TIGR Seventh Annual Conference on Computational Genomics*, Reston, VA, 21-24 October 2004.

TVFacDB -- A Comprehensive Microbial Toxins and Virulence Factors Database

Murray Wolinsky[1,*], Jian Song[1], Jason Gans[1], Cathy Cleland[1], Jingao Dong[1], Robert Leach[1], Chris Stubben[1], Yan Xu[1], Luther Lindler[2], Kevin Anderson[3], Frank Labeda[3], Elliot Lefkowitz[4], Alexander Morgan[5], Marc Colosimo[5], Alexander Yeh[5], and Lynette Hirschman[5]

[1]Los Alamos National Laboratory, Mail Stop M888, Los Alamos, New Mexico 87545; [2]Walter Reed Army Institute for Research; [3]US Army Medical Research Institute of Infectious Diseases, Ft. Detrick, MD 21702-5011; [4]University of Alabama at Birmingham, Birmingham, AL 35233; and [5]The MITRE Corporation MS K325, 202 Burlington Road, Bedford, MA 01730

In this joint effort with the University of Alabama at Birmingham, Walter Reed, MITRE and USAMRIID, we are developing a comprehensive database for microbial toxins and virulence factors (TVFac). This database will contain all known toxins and virulence factors as well as their homologs. Each TVFac record and the associated gene records will be fully annotated to include all relevant sequence and structural information as well as biological, epidemiological, and clinical information. Human annotation, sequence-based data analysis, protein structure prediction, and natural language-based data mining will be integrated for a comprehensive understanding of the common mechanisms leading to pathogenesis. This effort is intended to provide necessary infrastructure for studying the common mechanisms of virulence and identifying common "themes" utilized by microorganisms to cause disease. Bioinformatic tools are being developed to allow users to explore the database and perform detailed analysis and to support detection and countermeasure development efforts.

---

* To whom correspondence should be addressed

INTRODUCTION

Sequence-related research in biothreat reduction has focused on two broad goals: signature development to support detection and forensics, and understanding virulence factors to support countermeasures and treatment. Although both areas of research are important, it is our sense that current efforts in the biothreat community are not balanced between these goals but rather favor signature development. This imbalance is due to a prevailing organism-specific orientation. We propose to partially redress this imbalance by providing a resource that supports both signature development and analysis of the mechanisms of virulence in an evenhanded manner.

Specifically, we propose to develop a comprehensive database of microbial toxins and virulence factors, together with associated analytical tools and products. This resource differs substantially from existing databases in its cross-pathogen orientation: the database and tools are organized around pathogenicity rather than organism.

While the contents of the proposed resource are novel, the structure is not. The proposed configuration of capabilities is essentially identical to that presented in the (second-generation) annotated sequence databases that Los Alamos National Laboratory provides to the Department of Energy and to the National Institutes of Health.

Such a cross-pathogen resource will help improve existing signatures by helping relate these signatures to the pathogenic properties of organisms (making the signatures more robust as new "near-neighbors" are identified and also less easy to "engineer" away). It will also allow novel generic or "cross-pathogen" signatures to be developed.

This resource will support the computational and experimental study of common themes and variations in microbial pathogenicity. By focusing on common properties, the resource will provide a foundation for uncovering the essential building blocks or "piece-parts" of pathogenicity. In so doing, it will support the development of more effective therapies and preventative measures. The cross-pathogen orientation of the resource also facilitates the development of novel tools as described below.

Quick and accurate diagnosis of the diseases caused by any biothreat agents or agents of public health concerns is critical for developing timely and effective responses against bioterrorist attacks. Our current efforts have been focused on detecting and responding to only a few well-known agents (e.g., *Bacillus anthracis*, *Yersinia pestis*, *Clostridium botulinum*, smallpox, *etc.*). This focus must be broadened to include other agents, including novel (engineered) ones.

Better understanding of toxins and virulence factors is not only important for development of counter measures but is also essential for accurate clinical diagnosis. A great number of toxins and virulence factors are structurally and functionally related and share conserved domains or signatures. Any recombinant toxin and/or virulence factor (henceforth abbreviated TVFac) or new TVFac produced by other less known pathogens will likely share some of the same conserved domains or signatures that have been found in a known TVFac. So understanding of the common features shared by all of the known TVFacs will help us in detecting and responding to new or engineered biothreat agents.

All of our current efforts have been species-oriented and focused on identifying species-specific signatures. We have also devoted considerable effort towards development of many specific genome sequence databases to facilitate these species-oriented efforts (http://www.cbnp.lanl.gov/ and http://www.stdgen.lanl.gov/). However, there are few, if any, cross-species efforts made for the systemic study of microbial toxins. We believe a comprehensive microbial toxin database will provide the foundation needed for further computational and experimental analyses. Such database efforts will help us not only to develop methods for detecting a broad

range of biothreat agents, but also to build counter-measures that may be effective against many biothreat agents that produce similar toxins and virulence factors.

BACKGROUND

Pathogens must attach to host tissues, grow within these tissues while avoiding host defense mechanisms, and cause harm to their host(s). All of these activities exhibit common themes. For example, attachment to host tissues involves cell surface hydrophobicity; surface adhesins (pili); outer membrane proteins; glycoproteins; and/or enzymes which modify the host cell surface. Growth within host tissues involves iron (and other metal ion) acquisition systems; and acquisition or synthesis of essential nutrients (e.g., ABC transporters). Avoiding host defenses can involve any of a wide array of mechanisms including molecules which disable the immune system (cytokine mimetics); molecules which disable phagocytic cells; toxins which kill phagocytic cells; proteases which degrade antibodies; surface molecules which bind host molecules; and surface capsules which resist uptake by macrophages. Damage to the host can be created by assorted types of toxins, including enterotoxins, neurotoxins, membrane-damaging toxins, superantigens, etc.

Toxins and virulence factors are collectively these features of pathogens that interact with host cells to cause infection. In the absence of these factors, a pathogen may cause a milder form of disease, or no disease at all. Toxins are relatively unproblematic. However, it may be difficult to make a principled distinction between virulence factors and other components (e.g., "housekeeping genes") of pathogens involved in the normal activities of life. Nevertheless, we believe that there is adequate coherence to the concept of virulence factor and enough value to signature and counter measure development, to warrant an effort to catalog and annotate these factors. In particular, as our current resources do not allow us to address all factors simultaneously, we are starting with the least ambiguous cases and hope thereby to establish this resource on a sound footing.

OBJECTIVES

This database is being developed to serve the following objectives: (1) *Understanding virulence*: Virulence is a collective state of different virulence factors. Different virulence factors play different roles but all contribute to overall virulence. The unique set of virulence factors for a given genome determines what kind of infection it can cause, in what hosts, the types of tissue or cells it can infect, and its level of virulence. Studying the correlation between the types of infection and presence of a specific set of virulence factors will help us understand the roles each type of virulence factor plays during the infection process. (2) *Identification of the basic building blocks of virulence*: Different pathogens infect different hosts and cause different diseases due to the different sets of virulence factors they possess. This database will collect all known virulence factors and allow us to identify the basic building blocks of virulence. Identification and further understanding of such basic building blocks will help us to better understand the basic components of microbial virulence. (3) *Facilitating better detection of newly emerging or engineered biothreat agents*: Traditionally, the cause of infection or disease is determined by combination of organism identification using cultural or PCR methods and serological methods for the presence of specific antibodies against the toxins in the blood of the infected host. If recombinant agents are used, this traditional approach will fail or at best produce confusing results. Our analysis of related toxins will allow us to identify conserved domain structures. Short peptides that represent the conserved domains can be used to detect all related toxins. As more human pathogens are sequenced, the goal is to discover new toxins and virulence factors that will help us to better understand pathogenesis and to develop more effective counter-measures. In addition to their utility in diagnosis and vaccine development, the conserved domains can be translated into conserved corresponding DNA sequences. These conserved DNA sequences can be used as probes to actively search for new toxins in other pathogenic microbes. Even recombinant toxins will share the same conserved functional domains as other known toxins. With the help of this database, we may be able to reconstruct the creation of engineered pathogens by understanding which known toxins were used and how the recombinant toxins were created. (4) *Development of common strategies to counter any biothreat agents*: Similarly, the conserved domains identified can be used to induce protective immunity against a wide spectrum of pathogens that produce similar toxins; and (5) *Community resource*: because this database will contain a comprehensive collection of the known toxins, each of which will be thoroughly annotated, it will become a unique resource for many different end-users such as medical doctors, clinicians, experimental scientists, forensic scientists to meet their needs for background information. For example, clinicians will be able to find out which toxins cause what symptoms; which pathogens produces which toxins; what is known about the toxins – target tissues and the mode of action, and what counter-measures are available, *etc*. Experimental scientists can use the database to identify the conserved functional motifs present in new toxins for structure-function studies and can also find what has already been published on a particular toxin. Forensics scientists can use the database to find what toxins will target the tissue under examination. With the help of a broad interested community, we hope that the database develops into a unique and comprehensive resource supporting a variety of investigations.

RESULTS

Database Design and Implementation

Deficiencies in our scientific knowledge and a paucity of experts will ultimately limit our capability to rapidly and precisely identify agents and respond effectively in a crisis. For example, the global molecular epidemiology of the agents at the top of the threat list is critically important for identifying the organisms accurately and differentiating local from exotic strains. Current databases are inadequate, and no organized effort is being made to fill in

the gaps.
Phillip K. Russell, 1997.

Our initial design for this database has evolved considerably since its inception. The original concept of the database viewed it as a complement to existing annotated pathogen sequence databases. Consequently the original design stressed the novel, functional orientation the TVFac database is intended to offer. However, it soon became apparent that a number of investigations require the more traditional sequence oriented records as well as the newer, functional records. In particular, investigations into genomic islands (e.g., pathogenicity islands), as well as comparative studies of gene function in different phylogenetic groups of pathogens, require both sorts of information to be combined. Therefore, the current organization combines these record types. Our basic sequence oriented records contain information on:

- Genome alignments
- COGs, Pfam, BLOCKS, ProDom, PDB analyses
- Gene image map
- Functional class
- Pathways
- References
- Repeats
- rRNAs, tRNAs
- Structural features of proteins

as well as other data. These tables are shown in gray in Figure 1 below. The newer records describe features of organisms and denote such characteristics as:

- Pathogenesis
- Symptoms
- Detection
- Diagnosis
- Treatment

These records are shown in red in Figure 1. This organization allows us to support new investigations such as:

- Compare the neighboring regions of toxins and virulence factors
- Cluster genes in varying ways (i.e., based on sequence, functional and/or structural similarity)
- Load entire genomes for easy data mining and navigation
- Organize and manage TVFac hierarchies and relationships independent of gene records
- Inspect toxins and virulence factors under the context of pathogenicity islands
- Store and search organism-level notes on pathogenesis, symptoms, detection, etc.
- Store and search regulon information
- Search multiple fields from any table through the web interface
- Utilize repeat, IGS, tRNA and rRNA information

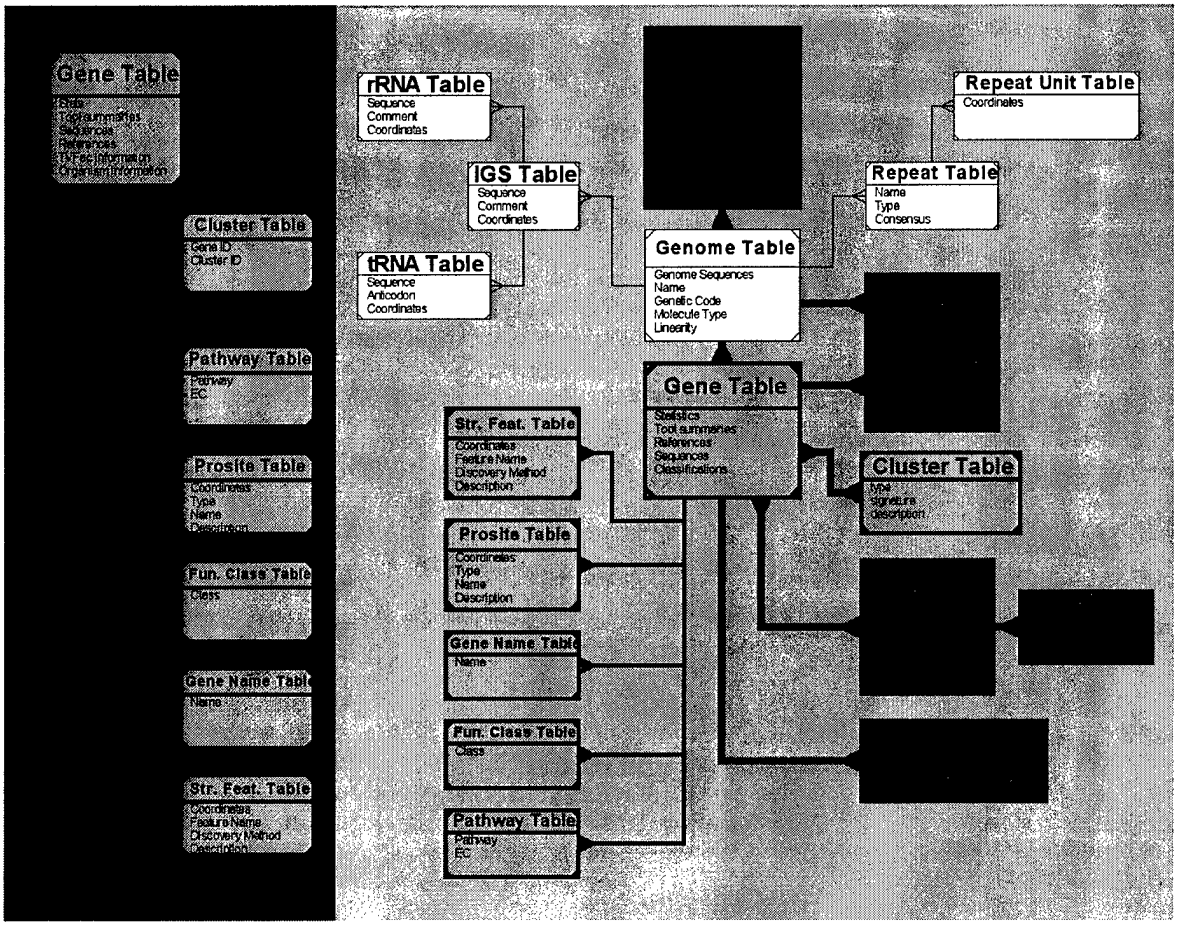


Figure 1. Table structure of TVFac database. Tables in gray contain sequence-oriented feature data. Those in red are new organism and "higher-order" data. Tables in white are primarily additional sequence-oriented data that has been added to the database.

Search and Browse capabilities

We are attempting, in conjunction with the MITRE Corporation, to produce a database that is much easier to use for both manual and automated services. To that end we are constructing a detailed set of use cases and designing the database around common sorts of questions that can be asked. There are two broad types of perspectives that users might possess. Users may take an organism-oriented point-of-view. From this perspective, questions of interest may be to find out what genes or properties a user-defined group of organisms has in common and what distinguishes the group of interest from other organisms. Another common perspective is a more functionally-oriented one. The user may not have in mind a particular set of organism; rather, he or she may be interested in a particular pathway, or functional feature, and want to know more information about that feature (including, possibly, which organisms share it). Our interface (see Figure 2) is designed to support both sets of queries. (The initial set of functional categories is shown in the appendix.)

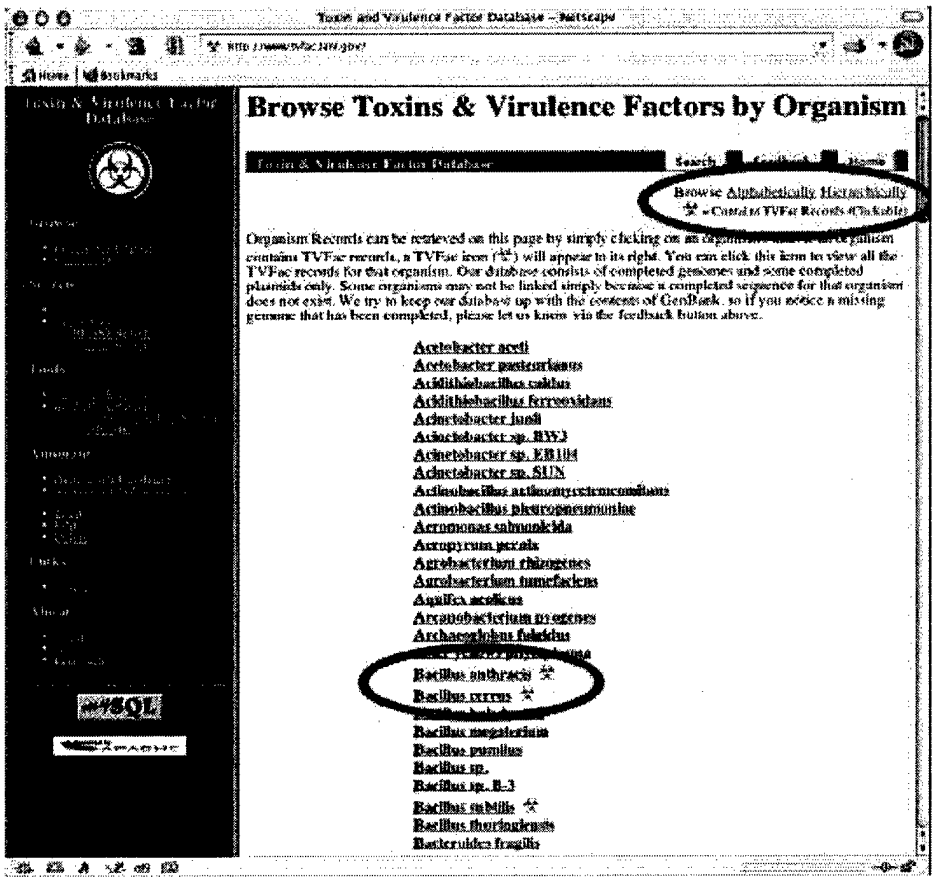


Figure 2. With help from MITRE, Los Alamos is attempting to provide a much friendlier interface. We support both an organism-centric perspective and a more functionally-oriented one for accessing and querying the database.

Populating the Database

Currently we are working on bacterial genomes, with plans to add viral genomes. All complete genomic sequences including chromosomal, plasmid, bacterial phages and the related annotation are downloaded from GenBank and used to initially populate our database. Each record in the database will be thoroughly annotated as in our genome sequence databases (http://www.cbnp.lanl.gov/ and http://www.stdgen.lanl.gov/). Annotation of a typical toxin record will contain more than 30 fields of information including basic characteristics of the genes belonging to a TVFac record (gene_name, gene_length, sequence) and the gene product (toxin) (length, MW, pI, net_charge, sequence), definition, functional_classification, EC number, cluster, Blast_summary, paralogs, Cluster of Orthologous Group (COGs), functional domain analysis (ProDom, Pfam, Prosite, Blocks), epitope prediction, structural features (Psort, PhD, SignalP predictions), PDB hits, primary and secondary references, etc.; other tables provide data regarding symptom, detection, target tissue, mode of action, and counter-measures (or prophylaxis).

In addition to full annotation, each record in the database will be clustered based on sequence similarity as well as structural similarity. Clusters derived from sequence and structure similarities will be related to function. We will build multiple alignments of related TVFacs and develop web interfaces to allow users to generate new alignments with any user-supplied sequences. Multiple alignments should allow us to identify conserved domains or signatures. We will also create Hidden Markov Models for each cluster of related toxins. These added capabilities will facilitate structural and functional studies and help build a better understanding of pathogenesis.

Bioinformatic Tools

The TVFac database is supplied with a collection of standard and novel bioinformatics tools. We have previously developed innovative genome viewers and phylogenetics browsers. For TVFac, we have integrated those tools, to provide a novel, integrated phylo-genomics viewer called NeighborView. (See Figure 3.) NeighborView allows users to visually compare the genomic neighborhoods of sets of organisms arranged phylogenetically. In a sense it presents a visualization of a "higher-order" sequence alignment, where it is individual annotated genes that are aligned, rather than their constituent nucleotides. This tool facilitates comparative analysis of function: in particular, it allows us to ascertain whether a specific gene appears to serve similar or distinct functions in differing groups of organisms.

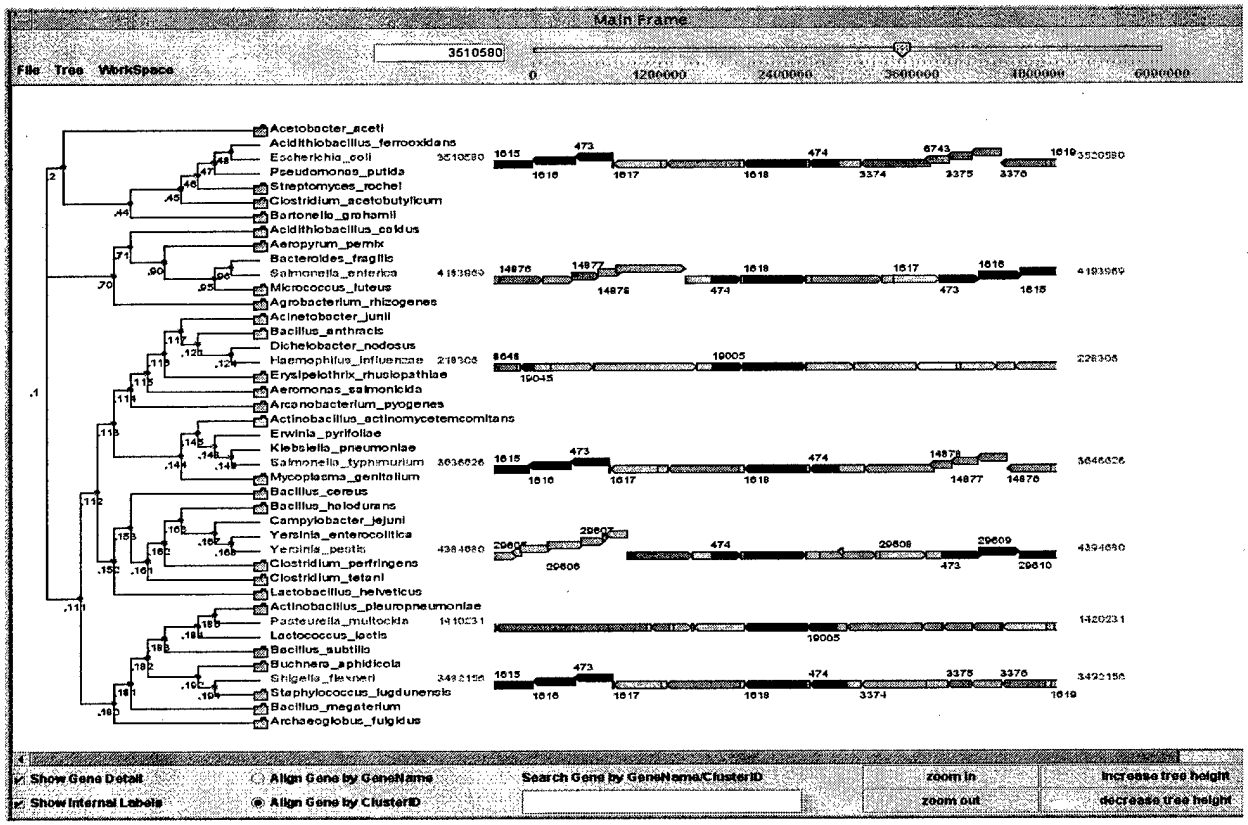


Figure 3. Screen capture of NeighborView tool. The region surrounding a specific gene is depicted for phylogenetically diverse organisms.

Another novel tool is provided by genWave, which is a tool designed to support investigations into locating genomic islands (e.g, pathogenicity islands). It provides a user-friendly interface to a wavelet based method of analyzing GC frequency changes which can indicate the presence of lateral gene transfer events. The method was originally suggested by Pietro *et al.*

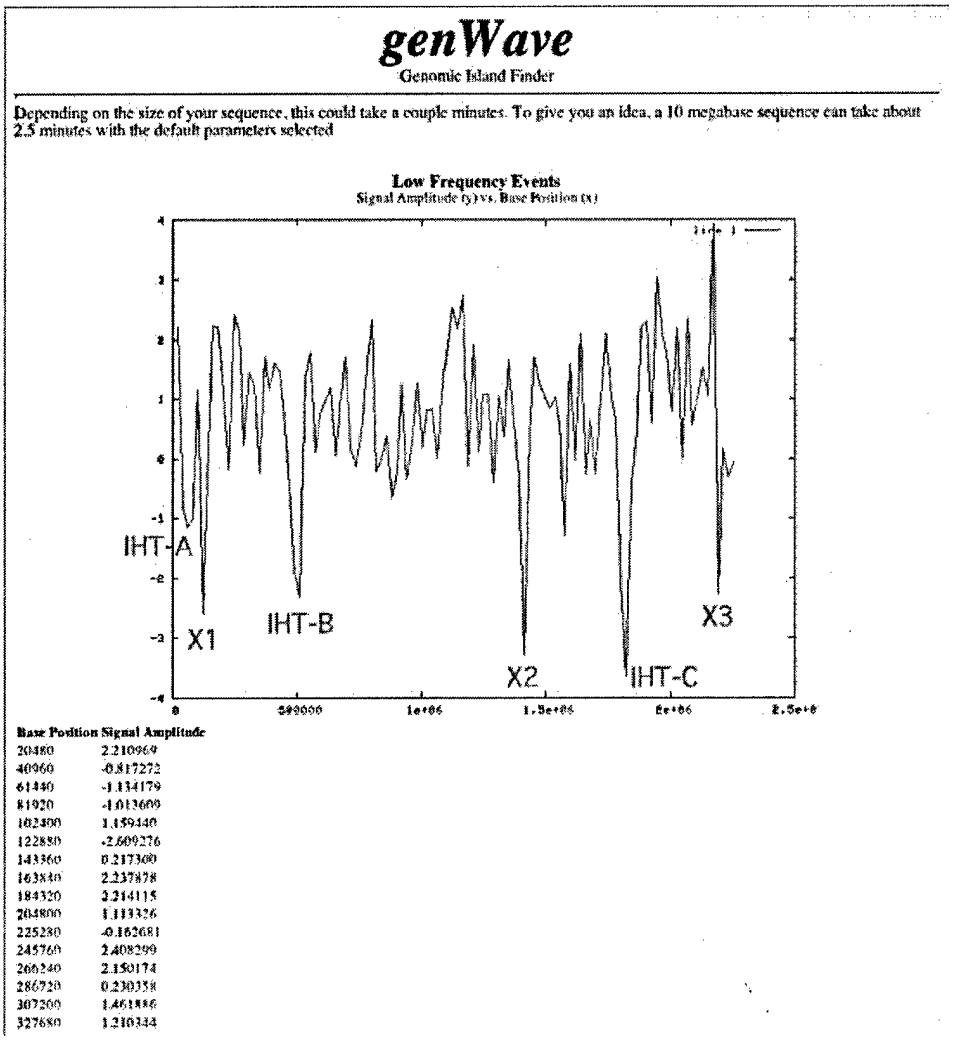


Figure 4. Replication of a study by Pietro et al using the genWave tool. Three known pathogenicity islands (labeled IHT-A, B and C are located and three more potential islands are predicted (labeled X1, X2, and X3.)

PRELIMINARY APPLICATIONS TO DEVELOPMENT OF GENERIC SIGNATURES AND COUNTER-MEASURES

We are actively trying to establish a user community for TVFac to help drive its early development. To that end, we are capitalizing on a shared interest with researchers at the Defence Science and Technology Laboratory (Dstl) at Porton Down, United Kingdom, in developing *generic* signatures and countermeasures for pathogens. Rather than detecting or treating a single group of related pathogens at a time, it would be ideal to be able to detect and treat broader classes of pathogens that share common mechanisms of virulence. This work is in its earliest stages and it remains unknown to what extent such generic signatures and countermeasures exist and have value. However, the goal of finding such generic features is interesting and worth pursuing. Moreover, our initial investigations have already pointed to some glaring deficiencies in existing knowledge and resources.

For example, we have done some preliminary work on identifying nucleic acid signatures for generic pathogens. We started by looking for nucleic acid fragments which are present in multiple pathogens, but which are not present in non-pathogens. Nucleic acid sequence data is of course readily available. The problem that one faces is identifying which sequences correspond to pathogenic organisms and which do not. *Surprisingly, this information is not readily available.* Therefore, in order to construct these (preliminary) generic signatures, we constructed an auxiliary database of pathogenic sequences – the pathogen list – which is part of the TVFac database.

The pathogen list

Constructing a list of pathogens is more problematic than might be apparent at first glance. Our initial work is intended merely to "move the ball forward:" additional insights and suggestions are more than welcome. Currently, our list of pathogens includes

representatives from bacteria, fungi, other eukaryotes, and viruses. These pathogens are associated with one or more hosts. We have used clinical definitions of pathogenicity as follows:

1. Pathogen - cause disease in some proportion of healthy individuals
2. Opportunist - cause disease if introduced into a protected site, or if normal host defense barriers are compromised
3. Rare opportunist - organisms known to cause infections in only a few clinical cases. (e.g. *Lactococcus lactis*, a commensal bacteria on cow udders with 7 known cases of human infections)
4. Non-pathogen - non-pathogenic strains in an otherwise pathogenic group (e.g., *E. coli* K12)

Other fields record the type of interaction, source, transmission, disease, references and any text supporting the pathogen classification. Currently, there are 580 species of bacteria in this list, which includes a preliminary list of 256 human bacterial pathogens and opportunists. Most of the pathogens in the pathogen table are classified at the species level, except when important differences in pathogenicity exist at lower levels in the taxonomic hierarchy. Types of interaction includes parasitic, commensal, zoonotic, and environmental. The names of diseases and links to external sources are also included.

A preliminary application of the TVFac database, its informatic tools, and the pathogen list, is shown in Figure 5, which depicts preliminary work in identifying potential nucleic acid signatures capable of distinguishing pathogens from non-pathogens. For a set of 77 pathogens (for which there is complete sequence information) we have been able to identify numerous sets of approximately 20 "minimal signatures" which reliably distinguish pathogens from non-pathogens. These signatures are comprised of nucleic acid 25-mers. These signature candidates were obtained using a rapid BLAST implementation, mpiBLAST, developed at Los Alamos, which runs on a dedicated 240 node Linux cluster. These nucleic acid stretches are candidates for generic signatures, and are potentially targets for creating generic countermeasures.

A second example of generic signatures, this time focused on a narrower group of pathogens, is provided by work in progress on a set of enteric bacteria of interest to a commercial partner. These organisms consist of *E. coli* CFT073, *E. coli* K12, *E. coli* O157H7, *E. coli* O157H7 EDL933, *S. typhi*, *S. typhi* Ty2, *S. typhimurium* LT2,*S. flexneri* 2a, *S. flexneri* 2a 2457T, *Y. pestis* CO92, *Y. pestis* KIM. Of particular interest are two fragments of length 40 which are present in all genomes of interest, but which are not present in any other organisms. One of these fragments comprises part of a specific gene and neighboring intergenic space. This region is shown in Figure 7 below. These fragments were identified as part of a more thorough analysis presented in Figure 6 below. This figure summarizes searches for DNA fragments of varying lengths which are common to at least three of the pathogens, but which are not present outside this set of organisms. There is no difficulty at all in obtaining thousands of fragments shared by most of these pathogens, but absent from all other organisms.

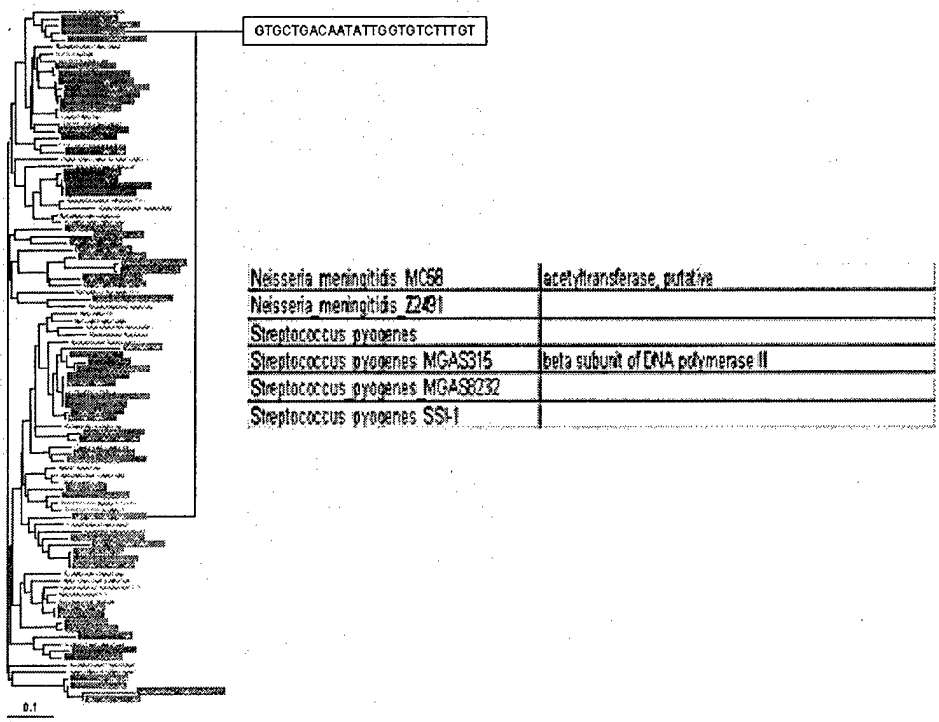


Figure 5. *This DNA 25-mer is present in a variety of pathogens (*Neisseria, Streptococcus *spp.) but is not present in all non-pathogens. We have found numerous sets of "minimal" generic signatures that can distinguish pathogens from non-pathogens. We have identified a set of 21 DNA 25-mers which appears to be sufficient to distinguish all pathogens from all non-pathogens.*

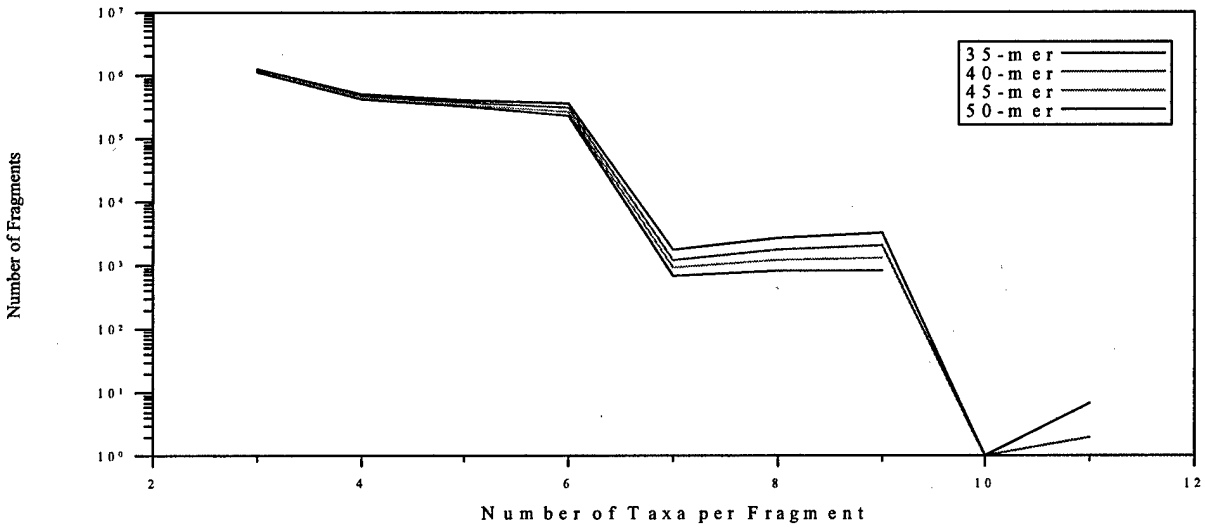


Figure 6. DNA fragments of varying lengths (from 35- to 50- mers) which are present in at least three of a set of enteric pathogens, but which are not present in any other organisms. We have found two 40-mers common to all 11 organisms of interest. These results were obtained on our 240 mode Linux cluster using a fast parallelization of the BLAST algorithm, mpiBLAST. (Note that the N-mers are distinguished by color and appear indistinguishable in black-and-white. ShorterN-mers appear above longer ones.)

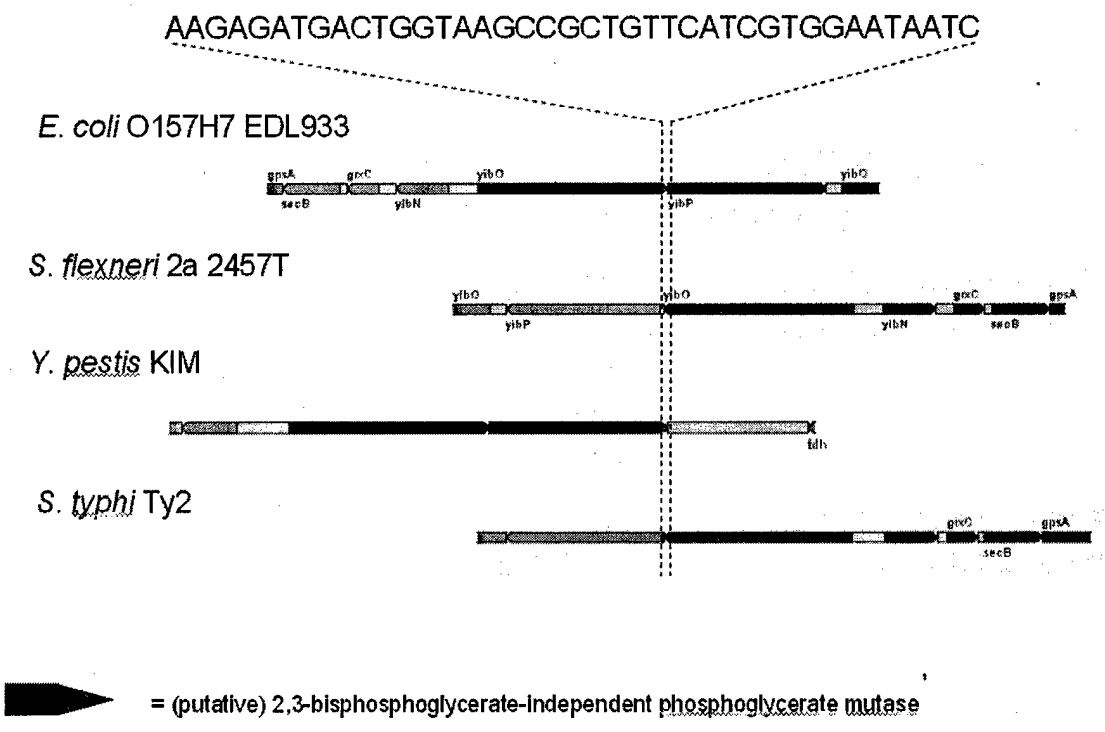


Figure 7.One of the two 40-mer fragments common to all target genomes

AVAILABLITY

TVFac is available to the public at http://www.TVFac.lanl.gov and most of the tools developed for the database are also be available to allow users to perform real-time and online analysis of their own genomic sequence data. The current version is mostly proof-of-concept and preliminary: much work is still required to make TVFac a useful resource.

FUTURE PLANS

Our current effort is to build a good infrastructure to allow expansion and community participation. In addition, we will add new types of data, including expression data and proteomic data as this data becomes available. We will further develop new comparative capabilities to allow us to predict virulence mechanisms and pathogenesis of new and emerging pathogens. Above all, we want to encourage interested users to participate in developing an invaluable resource for the entire biodefense community.

Appendix: TVFac Initial Functional Classifications

1. Exotoxins
   - 1.1. Others
   - 1.2. ADP-ribosylating toxins
   - 1.3. Proteolytic toxins
   - 1.4. Membrane-damaging toxins
     - 1.4.1. Others
     - 1.4.2. Phospholipases
     - 1.4.3. Pore-forming (channel-forming) toxins
     - 1.4.4. Membrane-disrupting toxins
     - 1.4.5. Hemolysins
     - 1.4.6. Leukotoxins
     - 1.4.7. Enteroaggregative exotoxins
   - 1.5. Superantigens
   - 1.6. Enterotoxins
   - 1.7. Neurotoxins
   - 1.8. A-B toxins
   - 1.9. RTX toxins
   - 1.10. Insecticidal
2. Adhesins
   - 2.1. Others
   - 2.2. P pili (fimbriae)
   - 2.3. Type IV pili
   - 2.4. Afimbrial adhesins
3. Invasins
   - 3.1. Others
   - 3.2. Hyaluronidase
   - 3.3. Collagenase
   - 3.4. Lecithinase
   - 3.5. Coagulase
   - 3.6. Fibrinolytic enzymes
   - 3.7. Flagella
   - 3.8. Induction of phagocytosis
4. Intracellular survival
   - 4.1. Others
   - 4.2. Capsules and LPS
   - 4.3. Oxidative stress responses
   - 4.4. Inhibition of oxidative burst
   - 4.5. Latency
   - 4.6. Prevention of phagolysome fusion
5. Anti-immune responses
   - 5.1. Others
   - 5.2. Antiphagocytosis
   - 5.3. Anti-complement
   - 5.4. Antigenic variation
   - 5.5. Phage variation
   - 5.6. Degradation of immune components
   - 5.7. Immunosuppression
6. Transport and secretion systems
   - 6.1. Others
   - 6.2. ABC transporter systems
   - 6.3. Type I secretion systems
   - 6.4. Type II secretion systems
   - 6.5. Type III secretion systems
   - 6.6. Type IV secretion systems
7. Endotoxins
   - 7.1. Others

7.2. Endotoxins
7.3. Endotoxin biosynthesis
8. Iron Acquisition
8.1. Others
8.2. Siderophores
8.3. Siderophore biosynthesis
8.4. Iron uptake
9. Antibiotics Resistance
9.1. Others
9.2. Antibiotics inactivation
9.3. Altered antibiotics target
9.4. Reduced antibiotics accumulation
9.5. Bypass antibiotics sensitive step
10. Regulation
10.1. Others
10.2. Transcriptional
10.2.1. Others
10.2.2. Two-component systems
10.2.3. Quorum-sensing
10.2.4. AraC faimily regulators
10.2.5. LysR family regulators
10.2.6. Fur proteins
10.2.7. Alternate Sigma factors
10.2.8. Anti-Sigma factors
10.3. Translational
10.4. Post-translational
11. Phage-related
12. Other viral
13. Unassigned

REFERENCES

Drell, S.D., A.D. Sofaer, and G.D. Wilson, *The new terror: facing the threat of biological and chemical weapons*. Stanford, CA. Hoover Institution Press, 1999.

Finlay, B.B. and S. Falkow, Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev*, 1997. 61(2): 136-69.

Kanehisa, M.I., Los Alamos sequence analysis package for nucleic acids and proteins. *Nucleic Acids Res*, 1982. 10(1): p. 183-96.

Xie, G., *et al*., Evolutionary origin of operonic genomic segments containing redundant copies of tryptophan-pathway genes: one in *Xyllella* species and the other in heterocystous cyanobacteria. *Genome Biol*, 2002. 4(R14).

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya, Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 2000. 288(5472): p. 1789-1796.

Fitch, J.P., et al., Biosignatures of pathogen and host. *Proc IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Raleigh, NC,, 2002.

Karlin, S., Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol*, 2001. 9(7): p. 335-43.

Yeh, A., L. Hirschman, A. Morgan, Evaluating text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, 2003. 19(Supplement 1): p. 331-339.

Hirschman, L., Morgan, A. and Yeh, A., Rutabaga by Any Other Name: Extracting Biological Names. *J. Biomed. Inform*., 2002. 35(4): p. 247-259.

Lebeda, F.J., Wolinsky, M. and Lefkowitz, E.J. "Information Resource and Database Development." In *Biological Weapons Defense: Principles and Mechanisms of Infectious Disease*, Lindler, L., Lebeda, F. J.and Korch, G., eds., Humana Press, Inc., Totowa, New Jersey (in press 2004).

Pietro *et al*. Finding Pathogenicity Islands and Gene Transfer Events in Genome Data, *Bioinformatics*, 2000

Riley, M., Functions of the gene products of *Escherichia coli*. *Microbiol Rev*, 1993. 57(4): p. 862-952.

Morgan, A., Hirschman, L., Yeh, A. and Colosimo, M. Gene Name Extraction Using FlyBase Resources. *ACL Workshop on Natural Language Processing in Biomedicine*. 2003. Sapporo, Japan.

Andrade, M.A., *et al*., Automated genome sequence analysis and annotation. *Bioinformatics*, 1999. 15(5): p. 391-412.

Forslund, D. Experiences Developing Distributed Object Applications. *OOPSLA* 1998. 1988. Vancouver, BC.

Feng, W.-c., The Design, Implementation, and Evaluation of mpiBLAST. *ClusterWorld*, 2003.

SNPSFINDER – *A WEB-BASED APPLICATION FOR GENOME-WIDE DISCOVERY OF SINGLE NUCLEOTIDE POLYMORPHISMS IN MICROBIAL GENOMES*

Jian Song[1], Yan Xu[1], Scott White[1], Kevin W. P. Miller[2], and Murray Wolinsky[1,*]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545 and [2]Counterterrorism and Forensic Science Research Unit, FBI Academy, Quantico, Virginia 22135

Running Title: *SNPsFinder*

* To whom correspondence should be addressed.

ABSTRACT

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variations among closely related microbial species, strains or isolates. Some SNPs confer selective advantages for microbial pathogens during infection and many others are powerful genetic markers for distinguishing closely related strains or isolates that could not be distinguished otherwise. To facilitate SNPs discovery in microbial genomes, we have developed a web-based application, *SNPsFinder*, for genome-wide identification of SNPs. *SNPsFinder* takes multiple genome sequences as input to identify SNPs within homologous regions. It can also take contig sequences and sequence quality scores from on-going sequencing projects for SNPs prediction. *SNPsFinder* will use genome sequence annotation if available and map the predicted SNPs regions to known genes or regions to assist further evaluation of the predicted SNPs for their functional significance. *SNPsFinder* can generate PCR primers for all predicted SNPs regions according to user's input parameters to facilitate experimental validation. The results from *SNPsFinder* analysis will be accessible through the World Wide Web.

Availability: The *SNPsFinder* program is available at http://snpsfinder.lanl.gov/.

Contact: jian@lanl.gov

Supplementary Information: The user's manual is available at http://snpsfinder.lanl.gov/UsersManual/

INTRODUCTION

Despite SNPs discovery has attracted much attention in human genome research, there are still relatively few studies on SNPs in microbial genomes. Most efforts have so far been focused on identifying unique genes or pathways in different microbial organisms through comparative genomic analysis. But the importance of SNPs in microbial genomes is being recognized. SNPs are the most abundant form of genetic variations in closely related genomes and the study of SNPs will undoubtedly offer new insights into many evolutionary processes including host-pathogen interaction (Blaser and Musser, 2001). In bacterial pathogens, a variety of SNPs have been discovered that confer a selective advantage during the course of a single infection, epidemic spread or long-term evolution of virulence (Ramaswamy *et al.*, 2003; Sokurenko *et al.*, 1999). SNPs contribute to the ability of pathogens to cause disease (Boddicker *et al.*, 2002; Weissman *et al.*, 2003). SNPs as genetic markers have been used to resolve closely related microbial species and

strains (Gutacker *et al.*, 2002) and to separate clinical samples collected from a disease outbreak facilitating investigations for infectious disease outbreaks (Cleland *et al.*, 2004; Read *et al.*, 2002).

The major limiting factor for SNPs discovery in microbial genomes has been the availability of the genome sequences. However, with high-throughput microbial genome sequencing projects worldwide, many closely related species and strains have recently been sequenced and many more are currently being sequenced. This is providing us unprecedented opportunity for genome-wide SNPs analysis. Here we report the development of the *SNPsFinder*, a web-based application for genome-wide SNPs discovery in microbial genomes. It takes multiple genomes as input and performs genome-wide SNP analysis. Using this application, we have successfully identified many useful SNPs that can be used as molecular signatures for clinical diagnostics and infectious disease surveillance and also help us to better understand the variations in both genotype and phenotype within many important pathogenic species.

ALGORITHM AND IMPLEMENTATION

Our goal is to develop a fully automated, genome-wide SNPs discovery program. To achieve this, we have developed an integrated algorithmic solution for the following five major tasks: (1) identifying all of the homologous regions among the multiple genomes being compared using MegaBlast (Zhang *et al.*, 2000); (2) eliminating paralogous sequences from consideration to reduce false positive SNPs identification; (3) generating multiple sequence alignments and detecting SNPs; (4) taking into consideration the quality of the sequences as well as the locations of the predicted SNPs to assist further evaluation of the predicted SNPs; and (5) picking up PCR primers for each predicted SNPs regions using Primer3 (Rozen and Skaletsky, 2000) to facilitate experimental validation. The major steps performed by *SNPsFinder* are summarized in Fig. 1 and a detailed description of algorithm and implementation is available in the supplementary information (User's Manual).

INPUT DATA

SNPsFinder allows the user to upload their genome sequences and other related data (genome annotation, sequence quality scores) from local files. The genome sequences can be either complete sequences or contig sequences. Users should choose a sequence that is in high quality and preferably has been annotated as the anchor sequence because the anchor sequence is used to map the predicted SNPs onto annotated genes or DNA regions to facilitate further evaluation of the predicted SNPs (Fig.1). When contig sequences are used, *SNPsFinder* allows inclusion of the corresponding sequence quality scores for consideration to reduce false positive SNP predictions as a result of sequencing errors. The user will be required to choose a percent sequence identity as a cut-off for *SNPsFinder* to determine what the homologous sequences will be identified and compared for SNP identification. The user is also required to choose a desired amplicon length (length of the homologous regions) by which anchor sequence will be fragmentized. To facilitate experimental validation of the predicted SNPs, *SNPsFinder* also allows the user to provide parameters according to which

PCR primers will be picked for each predicted SNP region. Finally, the user is required to enter an email address by which notification will be sent upon completion of the SNP analysis. A web link will be provided in the email to allow the user online access to the output of *SNPsFinder*.

OUTPUT OF *SNPsFinder*

The SNP regions identified by *SNPsFinder* are presented in a table format. They are sorted by the number of SNPs found in each region, but can also be sorted by genomic coordinates. The hyperlink for each predicted SNP region allows the user to view the corresponding multiple sequence alignment. Information on the gene that overlaps the predicted SNPs region is also provided and the gene IDs are linked to the GenBank records for more gene annotation. In addition to the SNPs, the numbers of insertions and deletions (InDels) found within the predicted SNP regions are also listed. Because the InDels within gene-coding regions often result in frame-shift and gene inactivation, thus identification of InDels will facilitate functional genomic analysis.

A pair of primers for each predicted SNP region is also generated by the *SNPsFinder* according to the parameters provided by the user. The user can view the primer information for a set of selected SNP regions or for all of the predicated SNP regions. The primer data include the SNP region ID, primer sequences, melting temperature, length of the primers and the expected amplicon length. The SNP region IDs are linked to the multiple sequence alignment where locations of the primer pair are labeled.

ACKNOWLEDGEMENT

This research was supported in part by DOE/DHS Chemical Biological National Security Program (CBNP) and the FBI. This work was also supported by the DOD/USAMRMC Toxin and Virulence Factor Database Effort (MIPR 2MCTC32157) and can be accessed through the website for that effort (www.tvfac.lanl.gov). We want to thank Electra Sutton for help in creating the *SNPsFinder* logo.

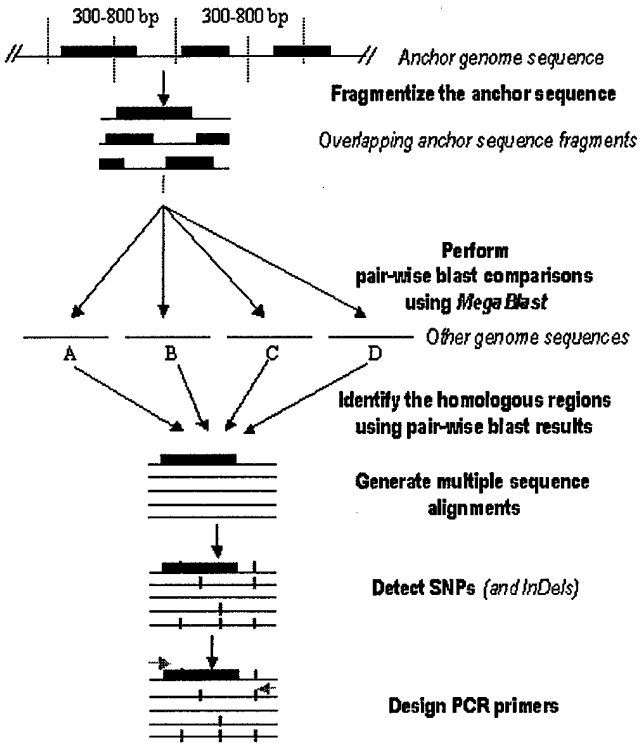


Fig. 1. Major steps performed by *SNPsFinder* for automated genome-wide SNP discovery. Solid rectangles indicate gene-coding regions and lines in between indicate intergenic regions. Short bars (red) indicate detected SNPs among the genomes in comparison and the arrows (green) indicate primer locations.

REFERENCES

Blaser, M.J. and Musser, J.M. (2001) Bacterial polymorphisms and disease in humans. *J Clin Invest*, **107**, 391-392.

Boddicker, J.D., Ledeboer, N.A., Jagnow, J., Jones, B.D. and Clegg, S. (2002) Differential binding to and biofilm formation on, HEp-2 cells by Salmonella enterica serovar Typhimurium is dependent upon allelic variation in the fimH gene of the fim gene cluster. *Mol Microbiol*, **45**, 1255-1265.

Cleland, C.A., White, P.S., Deshpande, A., Wolinsky, M., Song, J. and Nolan, J.P. (2004) Development of rationally designed nucleic acid signatures for microbial pathogens. *Expert Rev Mol Diagn*, **4**, 303-315.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N. and Musser, J.M. (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics*, **162**, 1533-1543.

Ramaswamy, S.V., Reich, R., Dou, S.J., Jasperse, L., Pan, X., Wanger, A., Quitugua, T. and Graviss, E.A. (2003) Single nucleotide polymorphisms in genes associated with isoniazid resistance in Mycobacterium tuberculosis. *Antimicrob Agents Chemother*, **47**, 1241-1250.

Read, T.D., Salzberg, S.L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J.D., Smith, K.L., Schupp, J.M., Solomon, D., Keim, P. and Fraser, C.M. (2002) Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science*, **296**, 2028-2033.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*, **132**, 365-386.

Sokurenko, E.V., Hasty, D.L. and Dykhuizen, D.E. (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol*, **7**, 191-195.

Weissman, S.J., Moseley, S.L., Dykhuizen, D.E. and Sokurenko, E.V. (2003) Enterobacterial adhesins and the case for studying SNPs in bacteria. *Trends Microbiol*, **11**, 115-117.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol*, **7**, 203-214.

PhyloNeighborView: a desktop application for gene-context comparisons against selectable 16S rRNA trees of microbial organisms

Yan Xu, Murray Wolinsky, Karla Atkins, and Jian Song*

Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Running Title: PhyloNeighborView

* To whom correspondence should be addressed.

ABSTRACT

Summary: PhyloNeighborView is a desktop application to assist biologists to explore completely sequenced microbial genomes and perform comparative genomics and phylogenomic analysis. It uses the complete genomes in TIGR Comprehensive Microbial Resource (CMR) (version 14.0) (Peterson *et al.*, 2001) as the genomic input and integrates the detailed, functional genomic annotation with a 16S rRNA-based phylogenetic tree to allow the viewing, comparison, and analysis of genes and their neighborhood across multiple species in evolutionary context. As a biologist-friendly tool, PhyloNeighborView requires little or no manipulation of genomic data by the user. It can be easily downloaded and run locally like any other desktop application.

Availability: PhyloNeighborView is available at http://biosphere.lanl.gov/PhyloNeighborView

Contact: yxu@lanl.gov or jian@lanl.gov

Supplementary Information: http://biosphere.lanl.gov/PhyloNeighborView/online_menu.html

INTRODUCTION

The ability to predict gene function based on sequence similarity has been of fundamental importance to progress in the interpretation of genomic data. However, sequence similarity alone often does not ensure identical functions (Eisen, 1998). Even with complete genome sequencing of many microbial species, there are still about half of the predicted genes that are of unknown biological function. Furthermore, it is common for a group of paralogous genes that are similar in sequence to have diverse functions as they have diverged after the gene duplication (Jensen *et al.*, 2002; Rison and Thornton, 2002). It has been estimated that the number of paralogous genes in a given genome increase from approximately 12-15% in genomes of 1Mbp up to approximately 50% in genomes of 3Mbp or larger (Fraser *et al.*, 2000). Thus, it is essential that we understand different functions of the paralogous genes in order to better understand the biology. But to do that, we will need to know how those genes have evolved and in what gene context (gene neighborhood) they reside among different genomes. Comparative genomic analysis with consideration of phylogenetic relationship will help us to improve the accuracy of predicting functions based only on sequence similarities (Eisen, 1998; Eisen and Wu, 2002). As more complete genomes become available, comparative analysis of multiple genomes will provide substantially more information on the physiology and evolution of microbial species and expand our ability to better assign putative function of uncharacterized genes (Fraser *et al.*, 2000).

Here we describe the development of PhyloNeighborView, a client-side, cross-platform pure Java visualization tool that integrates 16S-based phylogenetic tree with annotated genomes. We have integrated the completely sequenced genomes from TIGR CMR into PhyloNeighborView and we chose CMR because of its robust genome annotation. PhyloNeighborView allows visualization of gene-level comparisons across multiple species and uses the phylogenetic tree as a guide to display the gene neighborhood and analyze the level of gene conservation across tree nodes. Advantages of PhyloNeighborView over other software are (1) integration of phylogenetic tree to allow comparative genomics in phylogenetic context, (2) addition of sequence similarity-based clustering to allow users to browse multiple homologs (paralogs), (3) genes are color-coded according to their predicted biological functions, which will help to understand the functions of uncharacterized genes in the gene neighborhood, (4) biologist-friendly, no data manipulations are required and users simply download and run PhyloNeighborView locally like any other desktop application.

DESIGN GUIDELINES

Our goal is to develop a desktop application that (1) allows users to explore visually all of the complete microbial genomes; (2) is easy to deploy, install, update; and (3) is interactive, simple to use, and, most of all, is friendly to biologists. In order to achieve this goal, we implemented PhyloNeighborView as a client-side Java application and deploy the tool in our web server as a JNLP file link. Users can download the PhyloNeighborView from our website and run it with Java Web Start.

Users can launch PhyloNeighborView similarly as launching a native application. The requirement for client-side systems is to have Java 2 SDK version 1.4.x (where Java Web Start is a standard feature) installed.

Advantage of a client-side application over a web-based system is that no network connection is required and the application works locally once PhyloNeighborView is installed. Thus, the speed and performance of this application does not depend on the connection speed and there is no exchange of large chunks of data between client and server. Furthermore, integration of genomic data into the tool will save users from the needs of dealing with a huge amount of genomic data and performing data manipulations.

DISCRIPTION OF PHYLONEIGHBORVIEW

PhyloNeighborView can be launched from a user's desktop just like any other desktop application. PhyloNeighborView is a pure Java application and thus can execute on any platforms that have Java 1.4.x installed. PhyloNeighborView supports multiple workspace windows and has ability to save/reload a workspace to/from a file. Users can also print a current view or save it as a JPEG file.

PhyloNeighborView provides many ways for users to manipulate the phylogenetic tree and the gene images. For example, collapsing or expanding the tree nodes by simple mouse clicks, re-ordering sibling nodes, increasing/decreasing tree height, or locating a tree node from a sorted list of organism names. All of the gene images are clickable -- a single click to view the gene annotation and a double click to align genes across multiple selected genomes. Users can search and align a gene by either gene name or homology group and the target genes will be automatically aligned in the middle of the gene images. Users can also drag slider on top of the main window to move gene image map forward and backward and use the buttons on the bottom to zoom in/out the gene images. All genes are colored according to their functional classifications.

INPUT DATA

PhyloNeighborView uses two types of input data, a pre-computed phylogenetic tree based on 16S rRNA and abbreviated genome annotations (e.g., TIGR gene locus name, gi, gene start, gene stop, gene name, definition, main role and HG IDs). TIGR CMR (version 14.0) was downloaded from its ftp site (ftp://www.tigr.org/cmr/ftp/) and the 16S rRNA sequences and detailed genome annotation were parsed out. Sequence similarity-based clustering was performed using the E values from CMR all-against-all protein blast ('bcp_all_vs_all') as cutoffs to place all proteins into different homology groups (HG). Four separate clustering analyses, each using a different cutoff (e-50, e-25, e-10, or e-5) were performed. To make PhyloNeighborView a lightweight package for easy download and run, only abbreviated genome annotations and the 16S tree file are integrated with PhyloNeighborView. Detailed annotations were stored in a rational database and are accessible when 'Online' option is chosen.

GENERAL STEPS FOR USING PHYLONEIGHBORVIEW

i. Select and load genomic data for the species of interest by clicking organism names or by 'Loading genome data (in tree order)' from the 'Tree' menu (Fig. 1a).

ii. Double clicks on a gene image or enter a gene name into the text field below 'Search Gene by Name/HG' and press Return. All the genes with the same name will be aligned and displayed in the middle of gene images and labeled in red (Fig. 1b). Single click on any gene will display the gene annotation (Fig. 1c) and double clicks will result in re-alignment of the gene being clicked and its homologs across the selected genomes. If more than one gene with the same gene name are found within a genome, yellow arrow(s) on the far right side of the gene image will be displayed and the number associated with each arrow indicates the number of homologous genes and the direction of the arrow indicates whether the homologous genes are located in the upstream or downstream of the current highlighted (red) gene (Fig. 1b) and click on the arrow to locate the homologous genes.

iii. Homology group can be used just like gene name. It is, however, particularly useful in finding homologs within or across different genomes. Homologous genes often have different gene names and some of them may not even have gene names in the CMR annotation. By checking the radio button 'Or by Homology Group (HG) with Cutoff at' to switch to HG mode, HG IDs will be displayed around the gene images instead of the gene names. Under this mode, by double clicking on a gene image PhyloNeighborView will display all of the homologous genes that belong to the same homology group. Display of gene names or HG IDs can be switched back and forth by choosing either Gene Name mode or Homology Group mode.

iv. Finding a gene without gene name, either because it does not have one or its gene name is missing in TIGR CMR annotation, may not be as straightforward as finding a gene with the gene name in this current version, but can still be done. For example, chorismate mutase in *Agrobacterium tumefaciens* can be found through its neighboring gene *rpsP*. By finding *rpsP* first, the chorismate mutase gene can be located immediately upstream. Clicking on the chorismate mutase gene in *A. tumefaciens* under HG mode, users will be able to find all of the chorismate mutase genes if using a cutoff of e-10.

v. Finally, one can browse entire genome(s) by moving the slider on top the Main Frame window (Fig. 1b) or go to specific genomic location by entering a genome coordinate into the textbox in front of the slider.

ACKNOWLEDGEMENTS

This work was supported in part by DOE/DHS Chemical Biological National Security Program (CBNP) and the DOD/USAMRMC Toxin and Virulence Factor Database Effort (MIPR 2MCTC32157) and can be accessed through the website for that effort (www.tvfac.lanl.gov). We would like to thank Dr. Jason Gans for help in creating sequence-based clustering and Drs Roy Jensen and Gang Xie for valuable suggestions on improving PhyloNeighborView.

REFERENCES

Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, **8**, 163-167.

Eisen, J.A. and Wu, M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor Popul Biol*, **61**, 481-487.

Fraser, C.M., Eisen, J., Fleischmann, R.D., Ketchum, K.A. and Peterson, S. (2000) Comparative genomics and understanding of microbial biology. *Emerg Infect Dis*, **6**, 505-512.

Jensen, R.A., Xie, G., Calhoun, D.H. and Bonner, C.A. (2002) The correct phylogenetic relationship of KdsA (3-deoxy-d-manno-octulosonate 8-phosphate synthase) with one of two independently evolved classes of AroA (3-deoxy-d-arabino-heptulosonate 7-phosphate synthase). *J Mol Evol*, **54**, 416-423.

Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res*, **29**, 123-125.

Rison, S.C. and Thornton, J.M. (2002) Pathway evolution, structurally speaking. *Curr Opin Struct Biol*, **12**, 374-382.

Figure Legend

Fig. 1. Screenshots of PhyloNeighborView windows. **(a)** The Tree Selector window. A group of phylogeneticly related organisms can be selected by clicking a tree node ('Tree →Load genome data (in tree order) [*select your node(s)*]→Apply Selection'). **(b)** The Main Frame widow. On the left is the phylogenetic tree based on 16S rRNA, the organism names on the tree are clickable and can be used to load the corresponding genomes. On the right are the gene images with DNA molecule names (chromosome, plasmid, megaplasmid) to the left and genome coordinates on both ends. The start and stop coordinates of each genome region are colored in blue if the matched gene is on the same strand. Otherwise, the genes are flipped over and the coordinates are colored in green. The selected genes (*aroQ*) are center-aligned and colored in red and the surrounding genes are colored according to their functional classifications. **(c)** The Gene Annotation window. By single click on a gene image in the main window, the annotation window will pop out over the main window. By default, an 'offline' window is generated by PhyloNeighborView, which contains only abbreviated gene annotation and without any hyperlinks. However, by choosing 'Online' in the control panel, a dynamic window that contains detailed gene annotations with relevant hyperlinks will be generated remotely from our database.

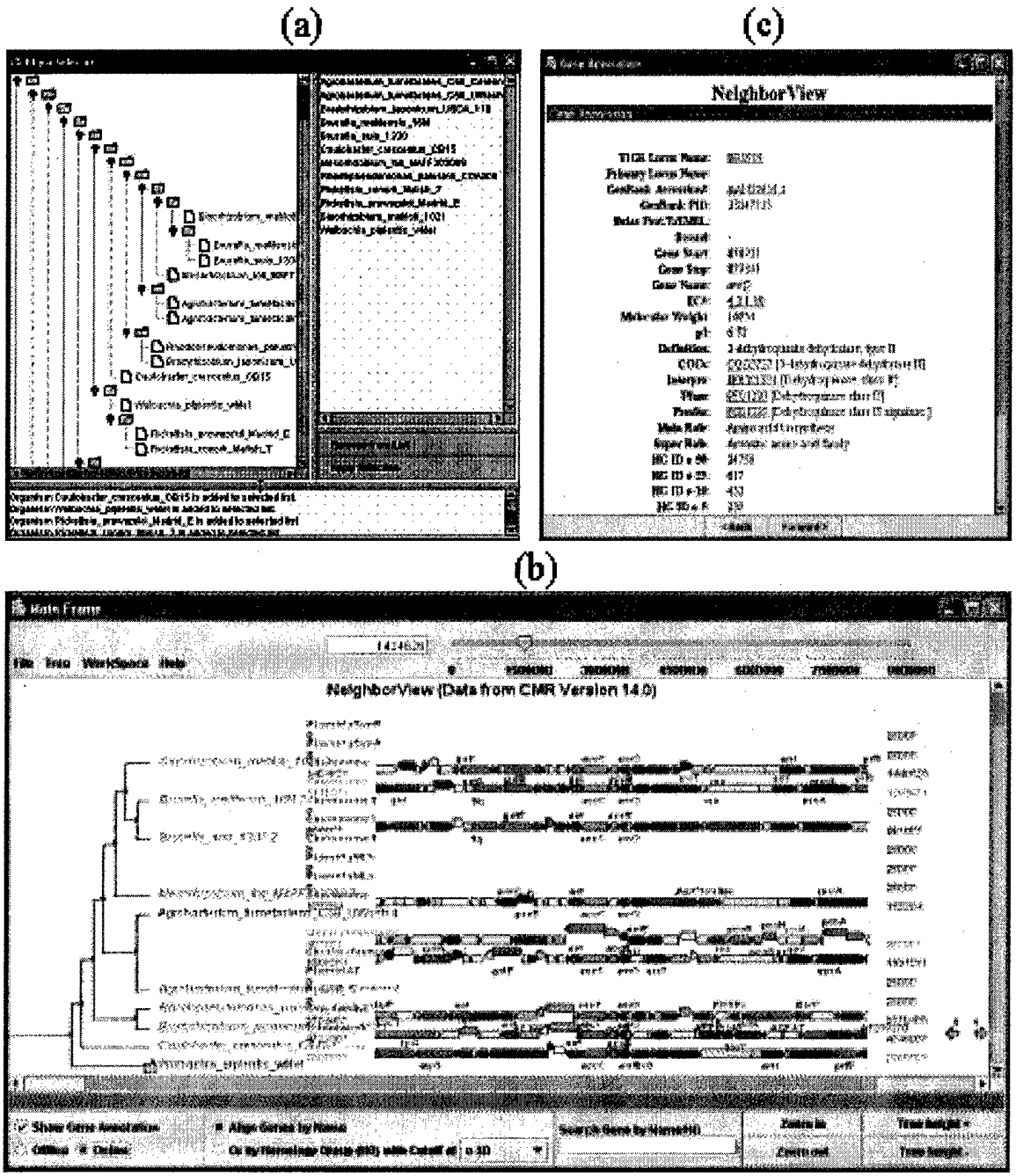


Fig. 1, Xu *et al.*

THE TYRA FAMILY OF AROMATIC-PATHWAY DEHYDROGENASES IN PHYLOGENETIC CONTEXT

Jian Song*[1], Carol A Bonner[2], Murray Wolinsky[1], Roy A Jensen [1,2]

Address: [1]Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA and

[2]Department of Microbiology & Cell Science, University of Florida, PO Box

110700, Gainesville, Florida, 32611, USA

Email addresses

Jian Song	jian@lanl.gov
Carol A Bonner	cbonner@ufl.edu
Murray Wolinsky	murray@lanl.gov
Roy A Jensen	rjensen@ufl.edu

*Corresponding author

ABSTRACT

Background: The TyrA protein family primarily catalyzes the key dehydrogenase reaction of *L*-tyrosine biosynthesis. All family members share in common a catalytic core region of about 30 kDa where inhibitors operate competitively by acting as substrate mimicks. This protein family is illustrative of the many that are challenging for bioinformatic analysis because of relatively modest sequence conservation and small size. Evolutionary variation of TyrA homologs abounds with respect to alternative specificities for each of the two substrates, fusion with other catalytic domains, and fusion with allosteric domains for regulation.

Results: The phylogenetic relationships of TyrA domains were evaluated in a context of the combinatorial patterns of specificity for the two substrates, as well as the presence or absence of a variety of fusions. An interactive tool is provided for prediction of substrate specificity, a feature not revealed by any obvious sequence motifs. An interactive alignment of core catalytic TyrA domains is also provided to facilitate phylogenetic analysis. *tyrA* membership in apparent operons (or supraoperons) was examined, and patterns of conserved synteny in relationship to organismal positions on the 16S rRNA tree were ascertained in *Bacteria*. A number of aromatic-pathway genes (*hisH*$_b$, *aroF*, *aroQ*) have fused with *tyrA*, and it must be more than coincidental that the free-standing counterparts of all of the latter fused genes exhibit a distinct trace of syntenic association.

Conclusion: We hypothesize that an ancestral TyrA dehydrogenase having broad specificity for both the cyclohexadienyl and pyridine nucleotide substrates existed. Indeed, TyrA proteins of this type persist today, but an abundance of examples also exists of narrowed substrate specificities, as well as of the acquisition of additional catalytic domains or regulatory domains via gene fusion. Particular domain fusions have occurred more than once. Fusions are stable markers of clades of differing hierarchical depth on phylogenetic trees. The evolutionary history of gene organizations that include *tyrA* can be deduced in genome assemblages of sufficiently close relatives, the most fruitful opportunities currently being in the Proteobacteria. The evolution of TyrA proteins within the broader context of how their regulation evolved and to what extent TyrA co-evolved with other genes as common members of aromatic-pathway regulons is an emerging topic of ongoing inquiry.

Dehydrogenases dedicated to *L*-tyrosine (TYR) biosynthesis comprise a family of TyrA homologs that have different specificities for the cyclohexadienyl substrate: ones that are specific for *L*-arogenate (AGN), specific for prephenate (PPA), or able to utilize both [1, 2]. Figure 1 illustrates the biochemical relationship of these specificities to divergent transformations beginning with chorismate utilization and converging upon TYR formation. This dehydrogenase protein family is equally diverse with respect to acceptance of the pyridine nucleotide co-substrate. Thus, a given TyrA enzyme having any of the aforementioned cyclohexadienyl specificities may be specific for NAD^+, for $NADP^+$, or may utilize either. This is consistent with a growing appreciation [3, 4] that different substrate specificities are often accommodated across a given protein family that nevertheless maintains a common scaffold of fundamental reaction chemistry. Even within the single category of broad TyrA specificity, there is a continuum ranging from examples where alternative substrates are accepted equally well to other cases where one substrate may be preferred by an order of magnitude or more. Table 1 provides a key to the nomenclature used to identify the various possible substrate-utilization combinations (both cyclohexadienyl and pyridine nucleotide) exhibited by TyrA proteins.

The TyrA family is typical of many protein families in that it consists of members having a relatively small core domain that is not highly conserved. As such, substantial challenges for bioinformatic analysis are posed. Here we have not only carried out a labor-intensive manual analysis, but we have developed some "template tools" that are intended to greatly facilitate and refine follow-on studies of this protein family.

RESULTS AND DISCUSSION

The TyrA phylogenetic tree

A large number of TyrA sequences (or TyrA domains in the case of fusions) were aligned, and this multiple alignment was used to obtain a protein tree (Fig. 2). The sources of TyrA proteins include all three domains of life, i.e., *Bacteria*, *Archaea*, and *Eukarya* (lower eukaryotes and higher plants). Figure 2 is intended to give a snapshot visualization of the overall complexity of the TyrA protein family from the vantage point of its multiple specificities, as well as its multiple fusion partners. A meaningful multiple alignment requires a trimmed set of sequences that correspond to the core catalytic domain. Alignment of sequences with non-homologous N-terminal fusions (such as with AroQ, $HisH_b$, or plant transit peptides) will appear to be more closely related on phylogenetic trees

than they are. This is because residues in the non-homologous N-terminal regions will find matches at random. Likewise, those sequences with C-terminal fusions (such as with AroF, ACT, or REG) will appear to be anomalously close.

Even enzyme proteins that have greater sequence conservation and amino-acid lengths than TyrA proteins cannot be expected to yield a protein tree that is congruent with the overall 16S rRNA tree. However, if genome representation is sufficiently dense within a range of closely related organisms, 16S rRNA-enzyme congruency can be expected within that change of organisms provided: (i) that the particular functional role has been retained and (ii) that lateral gene transfer has not occurred. This expectation follows from the outcome of a detailed analysis of tryptophan-pathway proteins in *Bacteria* [5, 6]. The majority of TyrA sequences available are from *Bacteria*, and one can see (by inspection of the major clades supported by high bootstrap values in Fig. 2) a general degree of congruence of the TyrA tree with 16S rRNA expectations of vertical genealogy. Thus, cyanobacteria, actinomycetes, low-GC gram-positive bacteria, and various divisions of the Proteobacteria separate into distinct cohesive clusters.

By far the most adequate genomic density available is for the proteobacteria. Figure 2 shows that alpha, beta and epsilon divisions of proteobacteria form cohesive clusters. Although delta-proteobacteria fall into two well-separated groupings denoted as delta_1 and delta_2, this should not be surprising since these groupings diverge at a deep level on the 16S rRNA tree, i.e., genome representation is poor in this region. On the other hand, genomic representation for the gamma-proteobacteria is excellent, but nevertheless their TyrA sequences separate into three well-spaced groupings, denoted gamma_1, gamma_2 and gamma_3 in Fig, 2. In this case, the separations seen between these fairly close relatives is attributed to particularly dynamic evolutionary events in the gamma proteobacteria (see later text). Tryptophan Congruency Groups 1 and 2 were based upon seven concatenated Trp-protein domains [6], and the gamma_1 and gamma_2 groupings of Fig. 2 correspond perfectly with these. One seeming exception is that the Trp-protein concatenates of *Xanthomonas* and *Xylella* were assigned to Tryptophan Congruency Group 2 [6], whereas the TyrA proteins of *Xanthomonas* and *Xyella* fall into the separate cluster (gamma_3) in Fig. 2. This is not especially surprising in that the *Xanthomonas/Xyella* lineage appears to be distinctly divergent from other gamma-proteobacteria [7], and indeed, Trp-protein concatenates from *Xanthomonas* and *Xyella*

were found to be outlying members of Tryptophan Congruency Group 2 [6]. Tyrosine congruency groups and tryptophan congruency groups are displayed in color code and mapped an 16S rRNA trees at **http://snp.lanl.gov/aroPath/TyrPath** and **http://snp.lanl.gov/AroPath/TrpPath**, respectively.

Distribution of TyrA specificity subclasses in Nature

Four qualitative classes of cyclohexadienyl substrate specificity populate the TyrA superfamily of homologs (Fig. 1). These include PPA-specific ($TyrA_p$), AGN-specific ($TyrA_a$), and the broad-specificity cyclohexadienyl ($TyrA_c$) dehydrogenases. A fourth class (Fig. 1) is represented by an enzyme of antibiotic biosynthesis (PapC) that converts 4-amino-4-deoxy-prephenate to 4-amino-phenylpyruvate [8]. Representatives of each specificity class have been studied at the molecular-genetic level. TyrA family members sharing a given substrate specificity do not necessarily cluster tightly together, and assignment of the substrate specificity of experimentally uncharacterized TyrA homologs is uncertain unless they exhibit very high amino acid identity with experimentally characterized TyrA proteins.

Cyclohexadienyl dehydrogenases. Most TyrA proteins (at least in the domain *Bacteria*) are of the $TyrA_c$ subclass. The cyclohexadienyl dehydrogenases commonly accept PPA or AGN about equally well, but various degrees of preference for one of the alternative substrates are also observed. Detailed molecular-genetic studies of $TyrA_c$ proteins from *Pseudomonas aeruginosa*, [9], *P. stutzeri* [1], and *Zymomonas mobilis* [10] have been carried out. A distinct variety of $TyrA_c$ merits separate status and has been denoted $TyrA_{c\_?}$. It exhibits a number of indels (mostly deletions) within the catalytic core region when its consensus sequence is aligned with those of the other TyrA classes (Fig. 3), and this correlates with the presence of an extra-core extension that may or may not have AroQ activity. Although the one large clade of $TyrA_{c\_?}$ proteins that has so far been studied prefers PPA over AGN by well over an order of magnitude, this does not reflect an obligatory relationship of preference for PPA and the presence of indels since a number of $TyrA_c$ proteins that lack the indels, e.g., $TyrA_c$ from *Neisseria gonorrhoeae*, also exhibit an overwhelming preference for PPA.

Arogenate dehydrogenases. The $TyrA_a$ class of specificity is currently represented by at least three widely spaced prokaryote lineages: cyanobacteria, coryneform bacteria, and *Nitrosomonas europaea*. This discontinuity of phylogenetic spacing is consistent with a basic evolutionary scenario [11] whereby the ancestral dehydrogenase was a broad-specificity $TyrA_c$ that evolved narrowed substrate specificity (to yield either $TyrA_p$ or $TyrA_a$) independently on multiple occasions in modern lineages. The high identities of TyrA sequences from *Mycobacterium tuberculosis*, *Bifidobacterium* (*Thermomonospora)* species, and perhaps *Streptomyces* species in Fig. 2 with that of *C. glutamicum* suggest a reasonable possibility that actinomycete bacteria as a group will prove to possess the same $TyrA_a$ specificity that is known to exist in coryneform bacteria. *Nitrosomonas europaea* currently

has no close genome relatives that have been sequenced. The first BLAST hit returned from a $_{NADP}TyrA_a$ query from *N. europaea* is the protein from *Ralstonia solanacearum*, which is known to differ from that of *N. europaea* with respect to specificity for both of its substrates (i.e., $_{NAD(P)}TyrA_c$) [12].

Recently, a plant *tyrA*$_a$ has been cloned and characterized from *Arabidopsis thaliana* [13]. Interestingly, the latter consists of two near-identical domains that are fused. The gene encoding this 68-kDa protein co-exists in the genome with a single-domain paralog [14] that encodes a predicted 37-kDa protein, somewhat larger than the core catalytic domain of $TyrA_a$ from *Synechocystis*. $TyrA_a$ (known to be located in higher-plant chloroplasts [2]) may have originated from cyanobacteria via endosymbiosis. If so, the plant $TyrA_a$ sequences have diverged sufficiently that they no longer share a specific phylogenetic grouping with the cyanobacterial TyrA sequences. This is in marked contrast with the phylogenetic proximity of the tryptophan synthase subunit proteins (TrpEa and TrpEb_1) from cyanobacteria and higher plants [15].

Prephenate dehydrogenases. $TyrA_p$ is most conspicuously represented by a large clade of gram-positive organisms, of which *Bacillus subtilis* $TyrA_p$ is the best studied [16]. At present, the latter comprise the only rigorously documented examples of absolute specificity for PPA. At the physiological level, those cyclohexadienyl dehydrogenases that exhibit a very substantial preference for prephenate are for all practical purposes prephenate dehydrogenases. These include the AroQ•$TyrA_{c\_?}$ enzymes of the enteric lineage (gamma_1 in Fig. 2). The $TyrA_c$ protein from *Neisseria gonorrhoeae* (and by inference, the closely related *N. meningitides*) is also a well-studied example of overwhelming preference for prephenate [12].

PapC dehydrogenases. PapC participates in the formation of *p*-aminophenylalanine as a step in the synthesis of at least two antibiotics (see Fig. 1). It is so far represented by only a few sequences. The PapC specificity is strongly indicated by absence of the otherwise invariant residue H197 (*E. coli* numbering) that is associated with recognition of a 4-hydroxy moiety in the cyclohexadienyl substrates of the aforementioned dehydrogenases, a moiety that is a 4-amino substituent for PapC dehydrogenase (see [17]).

The "redundant" *trp/aro* supraoperon of *Nostoc/Anabaena*

All cyanobacteria possess a highly conserved *tyrA*$_a$ gene, as well as a complete suite of tryptophan-pathway genes that are dispersed (unlinked) in the genome. The large-genome cyanobacterial lineage consisting of the *Nostoc* and *Anabaena* genera possess in addition a unique *trp/aro* supraoperon consisting of most of the aforementioned genes [18]. These include a second *tyrA* gene (denoted

tyrA$_{c\_?}$), six *trp*-pathway genes (all except *trpC*), and genes encoding the first two common-pathway steps of aromatic amino acid biosynthesis. All of these linked genes are seemingly redundant in that they are represented by homologs elsewhere in the *Nostoc* and *Anabaena* genomes at scattered loci. The closest BLAST hits for the *Nostoc/Anabaena* TyrA$_{c_?}$ proteins are not the co-existing TyrA$_a$ homologs that exist in their own genomes (and that are all universally present in all cyanobacteria). Rather the closest BLAST hits are to the TyrA$_{c_?}$ domains of the AroQ•TyrA$_{c_?}$ fusions in the enteric lineage. Since the enteric proteins are NAD^+-specific and strongly prefer prephenate, it is likely that the "extra" cyanobacterial proteins are also $_{NAD}$TyrA$_{c_?}$ proteins. Indeed, this would be consistent with enzymological evidence provided in the literature for both *Nostoc* and *Anabaena* [19].

Concerning the evolutionary origin of the redundant block of linked genes found in the *Nostoc* and *Anabaena* genomes, at least two possibilities await further illumination. (i) These genes might have been acquired by a common ancestor of *Nostoc* and *Anabaena* via lateral gene transfer. This is consistent with the observation that biosynthetic-pathway operons are generally absent in the cyanobacteria, and all of the linked genes could have been recruited in a single event. However, at present no candidate donor genomes are known that possess this supraoperon combination of genes. If the TyrA$_{c_?}$ proteins of *Nostoc/Anabaena* and of the enteric lineage are possibly related by LGT, it is of interest that the N-terminal extension of TyrA$_{c_?}$ from *Nostoc/Anabaena* resembles the AroQ domain of AroQ•TyrA$_{c_?}$ from enterics. In both cases the N-terminal residues may compensate for indel deletions within TyrA$_{c_?}$. Subsequently, AroQ function may have been acquired in one lineage (or lost in the other). (ii) Alternatively, *tyrA*$_a$ and *tyrA*$_{c\_?}$ (and the duplicated *trp* and *aro* genes present in the supraoperon) might be ancient paralogs within the cyanobacterial lineage. If so, at a time following divergence of heterocystous cyanobacteria from the unicellular cyanobacteria, the latter may have lost the clustered block of aromatic-pathway genes in a single event of reductive evolution. The supraoperonic genes might be related to a specialized function associated with "developmental" physiological processes that typify the filamentous, heterocyst-forming cyanobacteria. An illustrative precedent is the phenazine-pigment operon of *P. aeruginosa*, which combines unique phenazine-pathway genes with a redundant gene of common-pathway aromatic biosynthesis and two redundant (and fused) genes of tryptophan biosynthesis in order to accomplish the linkage of specific phenazine biosynthesis with a supply of 2-amino,2-deoxy-isochorismate, the branchpoint of divergence toward phenazine and tryptophan [20, 21]. This complexity is consistent with the large genome sizes of *Anabaena* (7.2 million bp) and *Nostoc* (9.2 million bp), compared with the much smaller unicellular genomes of *P. marinus* (1.7 million bp), *Synechococcus* (2.4 million bp), and *Synechocystis* (3.6 million bp).

Profile hidden Markov models (HMMs) to distinguish TyrA specificity subfamilies

The limited information thus far available about specific molecular roles of particular TyrA amino-acid residues has been summarized recently [17]. The core domains of known TyrA$_c$, TyrA$_a$, TyrA$_p$, and TyrA$_{c_?}$ proteins were aligned separately. No distinctive motifs were found that in isolation would be a clear predictive indicator of specificity for cyclohexadienyl substrate. Similar substrate specificity profiles probably can be dictated by alternative patterns of interplay between different residue combinations. The potential complexity is deepened by the likelihood that specificity determinants for cyclohexadienyl and pyridine nucleotide are not completely independent of one another.

Because of the rapid accumulation of incorrectly annotated TyrA entries in GenBank and other databases, partly due to complications of gene fusions and partly to failure to assimilate published substrate specificities, the use of BLAST does not return reliable results with respect to substrate specificity. Even the HMMs used in Pfam (**http://www.sanger.ac.uk/Software/Pfam/**) and Interpro (**http://www.ebi.ac.uk/interpro/**) were not helpful in this case because the HMM deployed in those databases was broadly but incorrectly defined as 'prephenate dehydrogenase ($NADP^+$) activity' (PF02153) for all TyrA dehydrogenases. However, Profile HMM is known to be well suited for modeling a particular sequence family of interest and for finding additional remote homologs [22]. It is reputed to outperform methods that rely only upon pair-wise alignment of homologous residues in predicting protein function [23]. Therefore, profile HMMs were constructed using our multiple sequence alignments of each curated TyrA subfamily, using the HMMER package [22].

The profile HMMs obtained are tentatively reliable for prediction of substrate specificity. To facilitate ongoing and future functional annotations, we have made our profile HMMs available as a resource for "specificity prediction" at **http://snp.lanl.gov/AroPath**/TyrPath. Users can match query sequences against the four profile HMMs to predict the subfamily to which a query sequence belongs. It is anticipated that future experimental data relevant to substrate specificity will facilitate refinement of the prediction program. For example, at present the program predicts that the TyrA sequences from organisms such as *Helicobacter pylori* and *Saccharomyces cerevisiae* belong to the $TyrA_a$ grouping, and it will be interesting to see whether this holds up to experimental confirmation. It is additionally fascinating that the dehydrogenase from *Archaeoglobus fulgidis* is predicted to belong to the indel-containing $TyrA_{c\_?}$ grouping and that it possesses an extra-core domain extension (an AroQ fusion), just as occurs for the large clade of enteric bacteria. However, the *Archaeoglobus aroQ* is fused at the C-terminal side of $TyrA_{c\_?}$, rather than at the N-terminus as is the case with enteric bacteria.

Users can enter query sequences into interactive multiple sequence alignments with any of the four catalytic-core seed sequences used for the profile HMMs. One can also align new query sequences with a selectable assemblage of the complete set of curator-approved TyrA catalytic-core TyrA sequences. In the latter case, any desired subset of seed sequences can be selected for alignment.

The core catalytic domain of TyrA proteins

The simplest set of TyrA proteins consists only of the core catalytic domain (about 180 amino acids) [1] and includes the well-characterized $TyrA_c$ enzymes from *Neisseria gonorrhoeae* [12], *Zymomonas mobilis* [10], and $TyrA_a$ from a cyanobacterium [17]. In addition the core catalytic domain from *P. stutzeri* has been engineered for study from a *tyrA*$_c$•*aroF* fusion [1]. These model core proteins do not cluster together on the TyrA protein tree (Fig. 2). Xie *et al.* [1] suggested that in this set of catalytic-core TyrA proteins, inhibitors bind at the catalytic site and exhibit classical competitive inhibition with respect to the particular cyclohexadienyl substrates that can be accepted by a given organism. This model predicts that the specificity for the sidechains of substrates utilized would parallel the specificity for inhibitor sidechains. The information summarized in Table 3 supports this expectation. Thus, the $TyrA_c$ proteins of *P. stutzeri* and *P. aeruginosa* will accept either a pyruvyl (as with PPA) or an alanyl (as with AGN) sidechain in the alternative substrates used, and this is paralleled by recognition of either a pyruvyl (4-hydroxyphenylpyruvate) or an alanyl (TYR) sidechain in the competent inhibitor structures. A variety of analog inhibitor structures were used by Xie *et al.* [1] to show that the minimal structure for binding at the substrate-binding site of *P. stutzeri* $TyrA_c$ is a six-membered ring with a 4-hydroxy substituent.

In contrast to the $TyrA_c$ proteins just described, the *Z. mobilis* $TyrA_c$ is totally insensitive to inhibition by either 4-hydroxyphenylpyruvate or TYR. Since both of these compounds lack a 1-carboxy moiety, it is reasonable to assume that the 1-carboxy substituent present in the two substrates accepted may be required for binding at the catalytic center. Thus, although $TyrA_c$ from *Z. mobilis* will accept the same two substrates as does the $TyrA_c$ from *P. stutzeri*, the greatly different inhibition results suggest that *Z. mobilis* obeys more stringent rules for binding at the catalytic site (i.e., a ring carboxylate must be present).

Synechocystis sp. and *A. thaliana* $TyrA_a$ proteins accept as a substrate only AGN, which has an alanyl sidechain. The ring-carboxylate moiety is evidently not absolutely required for binding since these $TyrA_a$ proteins can recognize TYR (alanyl sidechain) as an inhibitor. In contrast, since *N. europaea* $TyrA_a$ is not inhibited by TYR, it resembles the *Z. mobilis* $TyrA_c$ in the putative requirement for a 1-carboxy substituent for successful binding at the catalytic site. Finally, the *N. gonorrhoeae* $TyrA_{c\_?}$ exhibits an overwhelming substrate preference for PPA, and consistent with the foregoing, is subject to inhibition by 4-hydroxyphenylpyruvate but not by TYR.

In summary, some TyrA proteins probably require a 1-carboxy moiety for binding at the catalytic site, and these are insensitive to inhibition by the aromatic reaction products (which lack the 1-carboxy substituent). Other TyrA proteins not requiring the 1-carboxy moiety for binding will recognize product inhibitors that have the same sidechain as any substrate recognized.

Specificity for pyridine nucleotide cofactor within the TyrA superfamily

One can be fairly sure that TyrA proteins possessing D-32 (*E. coli* numbering) are NAD^+-specific. However, no obvious motifs could be found that distinguish TyrA proteins that are broadly specific for $NAD^+/NADP^+$ from those that are specific for $NADP^+$. As with the cyclohexadienyl co-substrate, narrowed specificity for NAD^+ (and $NADP^+$) also seems to have occurred independently on many occasions.

So far, the absolute specificity of $TyrA_p$ proteins for PPA is always accompanied by absolute specificity for NAD^+, although only a single large clade is thus far characterized. The opposite relationship, whereby absolute specificity for AGN is accompanied by absolute specificity for $NADP^+$, is also observed. Here three of the four $TyrA_a$ lineages described earlier exhibit this pattern. One exception, though, is the $TyrA_a$ of coryneform bacteria which accept either NAD^+ or $NADP^+$. Perhaps there is a general structural relationship that favors interaction between PPA and NAD^+, on the one hand, and between the greater positive charge of AGN and the greater negative charge of $NADP^+$, on the other hand?

$TyrA_c$ proteins tend to be NAD^+-specific, and this has been the property of the most rigorously characterized ones (from *Z. mobilis*, *P. stutzeri*, and *P. aeruginosa*). It is striking that in the pseudomonad clade marked by the *tyrA•aroF* fusion, the *Acinetobacter* sp. $TyrA_c$ is $NADP^+$-specific, whereas the sister subclade *Pseudomonas/Azotobacter* exhibits NAD^+ specificity (Fig. 2). Here the entire clade shares approximately the same profile of cyclohexadienyl substrate preference, but cofactor specificity has been narrowed in opposite directions.

Beyond the catalytic core: allosteric domains

Various lineages have acquired an amino acid binding domain known as the ACT domain (pfam01842), which is known to bind a variety of amino acids, thus functioning as allosteric domains for many proteins, including phosphoglycerate dehydrogenase, aspartokinase, acetolactate synthase, phenylalanine hydroxylase, prephenate dehydratase, and formyltetrahydrofolate deformylase. Recruitment of this domain by fusion with $tyrA_p$ appears to have occurred in a common ancestor of the large *Bacillus/Staphylococcus/Listeria/Streptococcus* assemblage (Fig. 2). It is interesting that *B. subtilis* also possesses a gene encoding a free-standing ACT domain in its genome (incorrectly annotated as *pheB*). Judging from the widely spaced tree positions (Fig. 2), additional fusions of

genes encoding an ACT domain and *tyrA* occurred in *Streptomyces coelicolor*, as well as in the common ancestor of *Xanthomonas* and *Xylella*. There is no correlation between presence of the ACT domain and specificity for cyclohexadienyl substrate since $TyrA_p$ from the *Bacillus* clade is PPA-specific, *Xanthomonas/Xylella* $TyrA_c$ is broadly specific, and *Streptomyces* TyrA is probably AGN-specific ($TyrA_a$).

B. subtilis, which belongs to the large clade having an ACT domain as a carboxy extension, has been extensively characterized [16]. 4-Hydroxyphenylpyruvate is an effective competitive inhibitor, as would be consistent with our proposed effects at the catalytic core for a PPA-specific enzyme. However, TYR, PHE and tryptophan were also inhibitors. The violation of the rule that the latter three amino acid inhibitors would not be expected to bind the catalytic core region (because they have alanyl sidechains even though the substrate-binding site only recognizes the pyruvyl sidechain of prephenate) and the finding that some of these were not competitive inhibitors can now be accounted for by the presence of the allosteric ACT domain. A carboxy extension shared by a number of *Archaea* (denoted 'REG' in Fig. 2) is presumably a regulatory domain as well. This is consistent with the recent result of Porat et *al.* [24] that not only 4-hydroxyphenylyruvate, but TYR, inhibited prephenate dehydrogenase activity of *Methanococcus maripaludis*.

The tyrA gene is a popular fusion partner

Fusion with *aroQ*. *tyrA* may be fused with a number of other catalytic domains, each of them relevant to aromatic biosynthesis (Fig. 2). *aroQ* is frequently fused with a number of other aromatic pathway genes [25]. The enteric lineage of *Bacteria* (defined as the wedge of gamma proteobacteria between *Shewanella putrefaciens* and *E. coli* on the 16S rRNA tree) possess an *aroQ•* $tyrA_{c\_?}$ fusion. The two protein domains of *aroQ•*$tyrA_{c\_?}$ may have co-evolved to produce cooperative protein-protein interactions since physical separation of the domains led to relatively low activities in *E. coli* [26]. The fusion physically links chorismate mutase (which forms PPA) with $TyrA_{c\_?}$ (which utilizes PPA). Exhaustive comparative enzymology has shown the *aroQ•*$tyrA_{c\_?}$ fusion to be stably maintained throughout the entire enteric lineage [27], except for instances of reductive evolutionary loss in pathogens (e.g., *Haemophilus ducreyi*) or endosymbionts (e.g., *Buchnera aphidicola*). An independent *aroQ•tyrA* fusion has occurred in the common ancestor of *Sulfolobus solfataricus* and *S. tokodaii* (Fig. 2). Since the TyrA domain of *Sulfolobus* species lacks the indel structure of the $TyrA_{c\_?}$ class, it would be interesting to see whether physical separation of the two domains would yield evidence of independent function, in contrast to the results mentioned just above for *E. coli*.

Fusion with *aroF*. Secondly, a cluster of pseudomonad organisms possesses a $tyrA_c$•*aroF* fusion. (*aroF* encodes enolpyruvylshikimate-3-P synthase, the sixth enzyme in the common pathway of aromatic biosynthesis; see [28, 29] for nomenclature used).This clade includes *P. aeruginosa*, *P.*

syringae, *P. putida*, *P. stutzeri*, *P. fluorescens* and *Azotobacter vinelandii*. It is interesting that *P. syringae* has experienced a deletion of about 200 residues at the N-terminal region of the AroF domain. This has been coupled with the acquisition of a monofunctional *aroF* gene that is absent in other members of the clade. Interestingly, the latter AroF shows high identity only with AroF from *Agrobacterium tumefaciens*, an alpha-proteobacterium. The *A. tumefaciens aroF*, in turn, is unique compared to its a-subdivision relatives, both in having divergent sequence and in being unlinked to *cmk* and *rpsA*. Thus, it seems likely that the incongruence of AroF belonging to both *P. syringae* and *A. tumefaciens* reflects acquisition via LGT from some as yet unknown source. The disruption of the fused *aroF* domain in *P. syringae* is an unusual instance where the function of one fusion domain has become discarded, yet the function of the second domain has been retained. An additional independent fusion of *tyrA* with *aroF* has occurred in the common ancestor of *Burkholderia pseudomallei* and *B. mallei*. This has been very recent since the closely related *B. fungorum* and *B. cepacia* organisms lack the fusion.

It has been suggested that presence of a given fusion may be useful for identification of clades that diverged from a common ancestor, independent of other methods [30]. Different fusions offer the power of discriminating clades at various hierarchical levels, i.e., nested clades discriminated by nested gene fusions. The *tyrA•aroF* fusion occurred in the common ancestor of the clade that includes the *Pseudomonas/Azotobacter* grouping and *Acinetobacter/Microbulbifer* (Fig. 2A). One can reasonably assume that relatively close Gram-negative organisms lacking the *tyrA•aroF* fusion diverged from the common ancestor of the fusion clade prior to the fusion event. It is interesting to note that TyrA within the *Acinetobacter/Microbulbifer* subgroup of TyrA•AroF proteins is $NADP^+$-specific, whereas the remaining pseudomonad group of TyrA•AroF proteins is NAD^+-specific. Thus, it is reasonable to conclude that the fusion event must have pre-dated the differential specialization for the pyridine nucleotide cofactor in the two sub-clades.

Fusion with *hisH*$_b$. Thirdly, a single organism, *Rhodobacter sphaeroides* possesses a *hisH*$_b$•*tyrA* fusion that must have occurred very recently. *hisH*$_b$ encodes an aromatic aminotransferase that is closely related to (or sometimes even synonymous with) imidazole acetol phosphate aminotransferase [31]. The *hisH*$_b$/*tyrA*/*aroF* linkage group is part of a supraoperon in some gram-negative bacteria in which a relatively conserved, yet frequently shuffled gene order is observed [28,

29]. Hence, it is reasonable to assume that at the time just prior to fusion, *hisH*$_b$, *tyrA* and *aroF* were adjacent. It would be interesting to know the substrate specificity of the *R. sphaeroides* TyrA domain. If it is AGN-specific the significance of *hisH*$_b$ presumably would be to transminate PPA to form AGN, the substrate used by TyrA$_a$ (see Fig. 1). On the other hand, if it is PPA-specific, the significance of the HisH$_b$ domain would be to transaminate the product of the TyrA$_p$ reaction. If the enzyme is a TyrA$_c$ enzyme, then HisH$_b$ likely is competent to catalyze either of the foregoing reactions.

Fusion with ACT. The widespread ACT regulatory domain appears to have been acquired by independent fusions at least three times judging from the widely separated lineages that possess a TyrA•ACT fusion (Fig. 2). Xie *et al.* [28] initially noted homologous domains positioned at the N-terminus of mammalian phenylalanine hydroxylase and at the C-terminus of most microbial prephenate dehydratases. This domain is responsible for phenylalanine-mediated activation and phenylalanine-mediated inhibition of the hydroxylase and dehydratase enzymes, respectively. This domain was later named the ACT domain [32] and shown to be a widely distributed domain family that shares a conserved overall fold. Members of the ACT-domain family possess a wide variety of different ligand-binding capabilities. For example, the ACT domain of 3-phosphoglycerate dehydrogenase binds *L*-serine as a allosteric inhibitor.

Fusion with REG. Another putative regulatory domain fused to tyrA and denoted *tyrA?REG* is thus far restricted to some of the *Archaea*. This domain is a predicted regulatory domain, as described in COG4937.

A novel 4-domain fusion. *Archaeoglobus fulgidus* exhibits a striking four-domain fusion consisting of three catalytic domains and a regulatory ACT domain (TyrA•AroQ•PheA•ACT).

***tyrA* in its syntenic context**

Although the genes of prokaryotes have clearly been subject to frequent scrambling, some gene-gene associations persist more tenaciously than others. Xie *et al.* [28, 29] asserted that one such ancestral gene string that has resisted scrambling forces is *hisH*$_b$ > *tyrA* > *aroF*. Another is *cmk* > *rpsA*. Gene synteny in prokaryotes has not been easily recognized in the past because substantial manual scrutiny in combination with a sufficient density of genomic representation on a given portion of the phylogenetic tree is necessary to detect patterns of synteny that are camouflaged by frequent scrambling events (inversion, deletion and transposition).

The domain *Bacteria* is now represented by a collection of sequenced genomes that is progressively approaching the genomic densities needed for meaningful analysis. Figure 4 provides a visual sense of the frequency with which *tyrA* is closely positioned with other genes of aromatic biosynthesis, as well as the underlying patterns of overall synteny. Even at a very deep level of phylogenetic branching, *Thermotoga* exhibits a *tyrA* gene flanked by seven genes encoding all of the common steps of aromatic biosynthesis. Although *tyrA* is not linked to any functionally relevant genes in *Aquifex*, representing the point of deepest phylogenetic branching, this does not necessarily mean that *tyrA* was not already generally associated with other aromatic-pathway genes at an early time. For reasons that are totally mysterious, certain scattered lineages exhibit a total lack of operon organization for aromatic-pathway genes

(and indeed for most other biosynthetic pathways, such as that for histidine biosynthesis). These lineages (Fig. 4) include, besides *Aquifex*, those of *Deinococcus*, the actinomycetes, the cyanobacteria, and *Chlorobium*. Except for the actinomycetes, this phenomenon of total gene dispersal also applies to genes of tryptophan biosynthesis [5, 6]. The *hisH*$_b$ > *tyrA* > *aroF* linkage first appears in *Bacillus* where an ancestral operon consists of *aroG* > *aroB* > *aroH* > *hisH*$_b$ > *tyrA*$_p$ > *aroF* [5]. *Bacillus* additionally possesses the *cmk* > *rpsA* unit in a separate location. Interestingly, in one narrow subclade (*B. subtilis*, *B. halodurans* and *B. stearothermophilus*) the *trp* operon has been inserted between *aroH* and *hisH*$_b$ to yield a supraoperon that has been fully characterized as a complex functional unit [33]. A pattern of association of *pheA* with *hisH*$_b$ > *tyrA* >*aroF* is suggested by linkage patterns in *Cytophaga* and *Bacteroides* (Fig. 4). *aroQ* became associated with *pheA* through gene fusion as early as the divergence of the *Spirochaetes* to yield an *aroQ•pheA*>*hisH*$_b$>*tyrA*>*aroF*>*cmk*>*rpsA* linkage unit (*Leptospira interrogans* in Fig. 4). The *aroQ•pheA* gene associated with *tyrA* and *aroF* in *Clostridium difficile* appears to have arisen from a different fusion event than that present in delta-, epsilon-, beta- and upper gamma-proteobacteria, on the one hand, or from that present in lower gamma-proteobacteria.

Detailed information that supports a deduced consensus for ancestral gene organizations with respect to beta proteobacteria, upper-gamma proteobacteria, and lower-gamma proteobacteria are shown later (Figs. 5-6). We suggest that the cenancestor of all proteobacteria possessed the gene organization *aroQ•pheA*>*hisH*$_b$>*tyrA*>*aroF*>*cmk*>*rpsA*. If so, the currently conserved *hisH*$_b$>*tyrA*>*aroF* and *cmk*>*rpsA* linkage units have since separated. The *aroQ•pheA*>*hisH*$_b$>*tyrA* portion likely specified all the catalytic requirements for conversion of chorismate to PHE and conversion of chorismate to TYR. Chorismate mutase activity specified by the *aroQ* domain could supply PPA for both PHE and TYR biosynthesis. Likewise, HisH$_b$, widely utilized as an aromatic aminotransferase [31], could also function for both PHE and TYR biosynthesis. Though currently available members of delta- and epsilon-proteobacteria exhibit substantial gene scrambling, the various fragmentary linkage patterns seen provide support for the ancestor proposed. *Geobacter* has the *aroQ•pheA* > *tyrA* > *aroF* > *cmk* > *rpsA* linkage group (with *lytB* inserted between *cmk* and *rpsA*). *Desulfovibrio vulgaris*, a delta-proteobacterium that is highly divergent from *Geobacter*, has a very interesting pattern of conservation and scrambling. *aroQ•pheA* > *aroF* > *tyrA* has been attached to a complete 7-gene *trp* operon. *hisH*$_b$ > *cmk* is completely separated from *rpsA*. The supraoperonic gene organization shown for *D. vulgaris* in Fig. 4 begins with two recently discovered gees, denoted *aroA'* and *aroB'*, that encode enzymes specifying an alternative biochemical route to dehydroquinate [34]. The epsilon-proteobacteria all display significant gene scrambling, but piecemeal evidence for the unscrambled ancestor proposed is present. For example, *C. jejuni* possesses an *aroQ•pheA* > *hisH*$_b$ unit, as well as *aroF* > *lytB* > *rpsA*. *Wollinella succinogenes* and *Helicobacter hepaticus* both possesses an *aroF* > *lytB* > *rpsA* unit.

The ancestor of alpha-proteobacteria appears to have lost the *aroQ•pheA* fusion and *pheA* is consistently monofunctional. Members of this group are quite uniform in the stable possession of *hisH*$_b$ > *tyrA* and *aroF* > *cmk* > *rpsA* as separated linkage groups. The beta proteobacteria are represented by members that have *serC* > *aroQ•pheA* > *hisH*$_b$ > *tyrA* > *aroF* > *cmk* > *rpsA*. This is also seen the members of the "upper" gamma-proteobacteria, except that *tyrA* and *aroF* are fused. The *serC* > *aroQ•pheA* > *hisH*$_b$ > *tyrA•aroF* > *cmk* > *rpsA* supraoperon has been studied in *P. stutzeri* [28, 29].

Zooming in on syntenic contexts of proteobacteria

Beta-proteobacteria and upper gamma-proteobacteria. The beta-proteobacteria exhibit a dynamic pattern of altered synteny (Fig. 5). Species of *Ralstonia* have retained the proposed ancestral synteny whereby the aromatic-gene unit *aroQ•pheA* > $hisH_b$ > *tyrA* > *aroF* is nested between *gyrA* > *serC* at the leftward flank and between *cmk* > *rpsA* > *himD* at the rightward flank. Species of *Burkholderia* (the next closest lineage) are almost identical, but exhibit individual events (as marked by circled numbers) of gene insertion, loss of $hisH_b$, dissociation from *rpsA* and *himD*, or gene fusion (fusion of *tyrA* and *aroF* in the common ancestor of *B. mallei* and *B. pseudomallei*).

At deeper levels in the tree, *Nitrosomonas europaea* exhibits a separation of the supraoperon between *tyrA* and *aroF*. Either a very large insertion was made between *tyrA* and *aroF* or one of these genes was transposed as a sufficiently large segment to include all of the conserved flanking genes. Species of *Neisseria* exhibit no remnants of supraoperon synteny at all, and the supraoperon genes have been completely dispersed. (It is interesting that among the beta-proteobacteria, *Neisseria* species are also unique in that all of the *trp*-pathway genes are dispersed [5]). In *Xylella* and *Xanthomonas*, $hisH_b$ has been deleted and *tyrA* has been transposed away from *serC* > *aroQ•pheA* < *aroF*. The latter unit has been transposed away from *gyrA*, the ancestral flanking gene. On the other hand, *cmk* > *rpsA* has remained next to *himD*, the gene usually flanking *rpsA*.

The gamma-proteobacteria have separated into two distinctly different synteny patterns. The *Pseudomonas/Azotobacter/Acinetobacter/Microbulbifer* assemblage shown in the lower part of Fig. 5 resemble the ß-proteobacteria. Like the *Ralstonia* grouping, an intact *ß*-proteobacteria-like supraoperon is present in *Pseudomonas aeruginosa* and *P. stutzeri*, differing only in the fusion of *tyrA* and aroF. The latter fusion occurred in the common ancestor of the upper gamma proteobacteria. *P. syringae, P. fluorescens* and *P. putida* have supraoperons lacking $hisH_b$. *P. syringae* exhibits a recent C-terminal truncation of *•aroF*, coupled with acquisition elsewhere in the genome of a free-standing *aroF* that is not phylogenetically congruent (probably of LGT origin). *Acinetobacter* sp. and *Microbulbifer degradans* possess an *aroQ•pheA* > *tyrA* > *aroF* unit that has become dissociated from the remaining supraoperon genes.

The enteric lineage. The lower-gamma proteobacteria, also denoted the enteric lineage by us, are shown in Fig. 3 as a clade of AroQ•TyrA fusions that exhibit absolute specificity for NAD^+ combined with an overwhelming but not complete specificity for PPA ($TyrA_{c\_?}$). In Fig. 6 the gene synteny of $tyrA_{c\_?}$ is profiled against the 16S rRNA phylogenetic trees of the gamma_1 proteobacteria possessing the genes shown. Fig. 4 has indicated a synteny consensus for the common ancestor in which *gyrA* > *serC* > $hisH_b$ > *aroF* > *cmk* > *rpsA* parallels the ancestral synteny of ß-proteobacteria, but without *aroQ•pheA* or *tyrA* in the middle of the linkage group. Many dynamic evolutionary events have intervened between the divergence of the lower- gamma proteobacteria from the upper-gamma proteobacteria. This includes the emergence of three allosterically distinct DAHP synthases, one of which now comprises the two-gene *tyr* operon ($aroA_{Ia\_Y}$ > $tyrA_{c\_?}$). The upper-gamma proteobacteria characteristically possess the $aroA_{Ia}$ homologs encoding $AroA_{Ia\_H}$ (Trp-inhibited) and $AroA_{Ia\_Y}$ (TYR-inhibited). It has been asserted that $AroA_{Ia\_F}$ was the most recent paralog, acquired just after divergence of the lower gamma-proteobacteria [35]. Therefore, the paralog conscripted into the *tyr* operon was already present in

the genome. The dissociation of $tyrA_{c\_?}$ from the *serC/rpsA* linkage group correlates with the fusion of *aroQ* with $tyrA_{c\_?}$. The *aroQ•pheA* has also escaped from the ser*C/rpsA* linkage grouping and has become linked with the newly emerged *tyr* operon. Some sort of duplication and recombinational event between *aroQ•pheA* and $tyrA_{c\_?}$ may have led to the creation of $aroQ•tyrA_{c\_?}$ since the AroQ•PheA proteins of gamma_1 proteobacteria are distinct from AroQ•PheA proteins of other proteobacteria with respect to the inter-domain linker length and the indel content (data not shown).

$HisH_b$ has persisted as the aromatic aminotransferase in the *Pasteurella/Haemophilus* grouping where two HisH paralogs are generally present, one of narrow specificity (denoted $HisH_n$) being within the histidine operon. The *aspC* gene next to *aroF* in *Shewanella* is a paralog that probably functions as an aromatic aminotransferase, much as *tyrB* in the *E. coli* grouping is a close paralog relative of *aspC* that has become specialized for aromatic biosynthesis [31]. Gene reduction associated with both endosymbiotic and pathogenic lifestyles are evident. Thus, *Buchnera* lacks *tyrA*, *cmk*, *hisH*, *tyrB*, and possesses only a single $aroA_{Ia}$ species ($aroA_{Ia\_H}$). *Haemophilus ducreyi* also lacks *tyrA*, as well as $aroA_{Ia\_H}$ and the entire *trp* operon [28].

TyrA in its context of regulation

TyrR regulon. Knowledge of the gene regulation exercised by TyrA in prokaryotes is sparse, being limited to the lower-gamma proteobacteria. Here, extensive information gathered from *E. coli* has revealed that $aroQ•tyrA_{c\_?}$ belongs to a large regulon controlled by the TyrR repressor. The limited phylogenetic distribution of TyrR, being present only in the lower-gamma proteobacteria (Fig. 7), indicates that it is a recent evolutionary acquisition. In *E. coli* the regulon members that are under the control of *tyrR* are the $aroA_{Ia\_F}$ > *tyrA* operon, the *aroLM* operon, *tyrP*, *tyrB*, *aroP*, *mtr*, $aroA_{Ia\_G}$, and *tyrR* itself [36]. Thus, *tyrR* not only regulates the tyrosine branch of the pathway, but heavily impacts the common pathway and the transport of all three aromatic amino acids as well.

Although outside the scope of this study, a logical expansion of it would be to examine the individual evolutionary histories of all the members of the contemporary *E. coli* tyrR regulon, i.e., asking when and in what order did these genes come under the influence of *tyrR*. Clearly, the recruitment of structural genes by *tyrR* has been recent, quite dynamic and even now, exhibits evidence of further ongoing change. For example, tyrosine phenol-lyase (a catabolic enzyme that is only sparsely present in gamma proteobacteria) has been recruited to the TyrR regulons of *Erwinia herbicola* [37] and *Citrobacter freundii* [38]. In these cases, not only does TyrR perform as a transcriptional activator, but it requires cyclic AMP receptor protein and integration host factor to do so.

As exemplified by *E. coli,* TyrR is generally a repressor. However, the transcriptional expression of *mtr* is activated by TyrR in the presence of TYR, and *tyrP* is activated in the presence of PHE (although it is repressed in the presence of TYR). The N-terminal domain of TyrR has been associated with the ability of TyrR to activate transcription in the case of *mtr* and *tyrP* [36]. The *Haemophilus/Pasteurella* lineage have all lost the N-terminal domain and presumably all lack the ability to accomplish transcriptional activation, as has been demonstrated experimentally with *H. influenzae* TyrR [39].

In view of the interesting complexity that two operons (*mtr* and *aroLM*) in *E. coli* are regulated by **both** *tyrR* and *trpR* [36], it may be more than coincidental that *tyrR* and *trpR* seen to have emerged at about the same evolutionary time, i.e., coincident with the

divergence of the upper-gamma proteobacteria from the lower-gamma proteobacteria (Fig. 6). A possible interaction between the TyrR and TrpR proteins has been noted [36].

PhhR in relationship to aromatic catabolism. Arias-Barrau et *al.* [40] have recently characterized a central catabolic pathway (Hmg) that degrades homogentisate in three steps as a source of carbon and energy to fumarate and acetoacetate. One of several peripheral pathways feeding into the central pathway begins with PHE and produces homogentisate via the reactions of phenylalanine hydroxylase (Phh), aromatic aminotransferase, and 4-hydroxyphenylpyruvate dioxygenase (Hpd). In the absence of Phh, a shorter version of the peripheral pathway is one that can use TYR, but not PHE, as a source of carbon and energy. In Fig. 7 the presence of Phh, Hpd, and Hmg segments of catabolism are mapped on a 16S rRNA tree. (The aromatic aminotransferase distribution is not shown since a multiplicity of aromatic aminotransferases having overlapping substrate specificities requires a particularly challenging effort in order to identify the functional role [31].) The cyanobacteria are unique among *Bacteria* in the utilization of Hpd for a completely different metabolic role that is unrelated to aromatic catabolism, i.e., the synthesis of vitamin E derivatives [41].

PhhR is a homolog of TyrR that has been shown in *P. aeruginosa* to be a divergently transcribed activator of a 3-gene operon that is needed for PHE and TYR catabolism [42]. The structural genes encode phenylalanine hydroxylase (*phhA*), carbinolamine dehydratase (*phhB*), and 4-hydroxyphenylpyruvate aminotransferase (*phhC*) and are powered by a s^{54} promoter [42, 43]. PhhR evolved relatively recently since it is only present in some gamma-proteobacteria (Fig. 7). The ancestral regulatory gene for the Phh peripheral pathway may have been a member of the leucine-responsive regulatory protein/asparagine synthase C (Lrp/AsnC) family judging from the adjacent and divergently oriented position of ASNC genes to *phhA* in organisms such as *Xanthomonas axonopodis* and *Mesorhizobium loti.* A recent overview of the many different regulator families involved in the control of aromatic catabolism conveys an emerging sense of the variety and dynamic evolutionary processes that underlie aromatic catabolism [44]. Occasional distant homologs of *phhR* that appear in erratic fashion in Fig. 7 may have some other regulatory function. For example, perhaps *Clostridium tetani* utilizes its *phhR* homolog as a transcriptional activator of the gene encoding tyrosine phenol-lyase, as occurs in species of *Erwinia* [37] and Citrobacter [38]?

Relationship of TyrR and PhhR. What might be of origin of TyrR? TyrR is an anomalous member of the large prokaryote family of s^{54} enhancer-binding proteins that activate promoters dependent upon s^{54}. Oddly, TyrR is the only homolog member that regulates s^{70} promoters, usually (but not always) being a repressor. Its closest homolog relative is PhhR, a canonical member of s^{54} enhancer-binding proteins. s^{54}-dependent enhancer proteins possess a highly conserved s^{54}-contact motif, GAFTGA, that is intimately involved in formation of the ternary complex of enhancer and s^{54}-RNA polymerase holoenzyme [45]. This is perfectly or nearly perfectly retained in the upper clades shown in Fig. 8, but is disrupted or completely absent in the clades between *Shewanella oneidensis* and *Pasteurella multocida* in Fig. 8. The deeper phylogenetic distribution of PhhR (Fig. 7) suggests that TyrR evolved as a variant of PhhR. If correct, a regulatory gene that is oriented to catabolism (*phhR*), and itself of relatively recent origin, was conscripted even more recently for a role in the regulation of primary biosynthesis (*tyrR*).

Consistent with the latter supposition, the gain of TyrR generally correlates with the loss of competence for aromatic catabolism (Fig. 7). In contrast to be *Citrobacter/Salmonella/Escherichia/Shigella* and the *Pasteurella/Haemophilus* clades (which completely lack the

GAFTGA motif), the remaining enteric clades have retained some residues in this region. These residues appear to be more than random remnants. It would be interesting to know if these residues have any functional significance. Indeed, the *Photobacterium/Vibrio* clade have retained the ancestral catabolic capabilities (Fig. 7) that would appear to demand retention of PhhR regulation; yet the parallelism of overall biosynthesis features with other lower-gamma proteobacteria would seem, on the other hand, to demand TyrR-mediated regulation. Perhaps this "TyrR" species participates in the regulation of both catabolic and biosynthetic genes. In this connection, it is interesting that Chaney *et al.* [45] found that a change in the GAFTGA motif of NifA could be partially "suppressed" by mutational changes in the N-terminal region of s^{54}.

Even more striking as a possible evolutionary intermediate is the most outlying member of the lower- gamma proteobacteria, *Shewanella oneidensis*. The position of TyrR on the protein tree parallels expectations based on the 16S rRNA tree. This, plus the conservation of the TyrR regulon features and the overall gene synteny suggest *E. coli*–like function as TyrR as a general repressor of regulon-member s^{70} promoters engaged in aromatic biosynthesis. However, the location of "TyrR" in *S. oneidensis* between PhhA and PhhB on one side, and *HmgB* and *HmgC* on the other side, imply some kind of regulatory relationship with the catabolic genes. It would be quite interesting to determine experimentally whether "TyrR" in *S. oneidensis* (and maybe *Vibrio*, as well) can function both as a repressor of the usual suite of s^{70} promoters, as well as an activator of promoters for phhA/phhB and/or hmgB/hmgC.

In view of the distribution of genes encoding catabolic enzymes, PhhR and TyrR, the most parsimonious evolutionary scenario may be that central and peripheral catabolic pathways depicted in Fig. 7 are quite ancient, but emergence of PhhR as a s^{54}-dependent activator of phenylalanine hydroxylase was quite recent, originating about the time of divergence of gamma-proteobacteria. The clade defined by *Shewanella/Vibrio/Photobacterium* retained the catabolic pathway, whereas the other enteric lineages discarded the catabolic pathway, but retained PhhR, which was then recruited as a s^{70}-dependent regulator of aromatic biosynthesis (TyrR).

Regulation by attenuation. A widespread mechanism of regulation is via an attenuation mechanism whereby transcripts that are initiated at given promoters can be terminated prior to reaching the structural genes of an operon. Whether termination occurs usually depends upon the balance between mutually exclusive terminator and anti-terminator structures, the balance being dictated by a variety of mechanisms [46].

Merino and Yanofsky have developed a website to provide a database of putative attenuators ahead of various operons in *Bacteria* (http://cmgm.stanford.edu/~merino/). We screened this database for likely attenuators relevant to the regulation of *tyrA*. Table 6 shows intriguing results that point to significant experimental work that would be desirable. *tyrA* is frequently a member of apparent supraoperons, as alluded to elsewhere in this paper, and some of these appear to be large gene clusters controlled by attenuation. Substantial work is needed to establish the depth of clades possessing a given attenuator. For example, the $hisH_b$ > *tyrA* operon is reliably present throughout all of the alpha-proteobacteria, *Agrobacterium tumefaciens* has been found to possess an attenuator, ahead of the $hisH_b$ > *tyrA* operon and one expects that most or all of the alpha-proteobacteria will possess the attenuator as well. If not, this

attenuator would appear to have been a very recent evolutionary innovation. Likewise, although the *aroA*$_{Ia\_Y}$ > *tyrA* operon is widely present throughout the lower gamma-proteobacteria, only several species of *Vibrio* appear to possess the attenuator.

Some of the supraoperons are very interesting in that they contain the majority of genes needed for both PHE and TYR biosynthesis, e.g., *Enterococcus faecalis* and *Streptococcus pneumoniae*. The latter organism displays two attenuator units. The supraoperon of *Desulfovibrio vulgaris* is novel in that it begins with two relatively rare genes encoding alternative enzyme steps for aromatic biosynthesis [34], denoted here as *aroA'* and *aroB'*. The leading five genes are adjacent to the seven-gene *trp* operon.

CONCLUSIONS

The evolutionary analysis of *trp*-pathway genes [5, 6] can be used as a template for comparable studies with other gene systems. Expansion to the greater aromatic pathway is a logical extension. The dynamics of evolutionary change for *tyrA* can be matched to the dynamics exhibited by the *trp* system. For example, beta-proteobacteria and "upper" gamma-proteobacteria separate as a distinct phylogenetic unit from "lower" gamma-proteobacteria with respect to dynamic milestone events that occurred with respect to many character states present in the lower gamma-proteobacteria. In the future one can anticipate that comprehensive and systematic phylogenetic analysis of each protein member of the TYR, PHE and TRP branches, the common aromatic-pathway trunk, and minor vitamin-like branches (such as the PABA/folate branch) will accommodate an integrated picture of the entire aromatic network, including catabolic pathways and many other specialized pathways.

METHODS

Alignments. **Multiple alignments were obtained by use of the ClustalW or ClustalX programs (Version 1.83) [47]. Manual adjustments were needed in the region of the GxGxxG motif for binding of pyridine nucleotide cofactor in the N-terminal region of TyrA proteins, and this was done with the assistance of the BioEdit multiple alignment tool of Hall (5.0.9 Edition) [http://www.mbio.ncsu.edu/BioEdit/bioedit.html]. The refined multiple alignment was used as input for generation of a phylogenetic tree using the phylogeny inference package (Version 3.2), PHYLIP [48]. The neighbor-joining program was used to obtain a distance-based tree. The distance matrix was obtained by use of Prodist with a Dayhoff Pam matrix. The Seqboot and Consense programs were then applied to assess the statistical support of the tree using bootstrap resampling (1,000 replications). We also used the ANCESCON package obtained from Cai et al. [49], and this produced similar results as given in Fig. 2, albeit with even wider separation of many groups. The presence of regulatory domains (ACT and REG) was accepted when indicated by the Domain Architecture Retrieval Tool (DART) on the BLAST menu at NCBI.**

Profile Hmms. Profile hidden Markov models for each of the four TyrA subfamilies, $TyrA_a$, $TyrA_c$, $TyrA_p$ and $tyrA_{c\_?}$, were built using Sean Eddy's HMMER package [50]. The bacterial species used for the seed sequences to generate the HMMs for each subfamilies can be found in Fig. 2, and only the core sequences that correspond the TyrA catalytic domains [17] are used. The seed sequences for each subfamily were first aligned using ClustalW [47] and the resulting multiple sequence alignments were then manually edited to produce more accurate alignment of the seed sequences, and finally the edited multiple sequence alignments were used to generate the profile HMMs for each TyrA subfamily.

Appraisal of gene fusions as one-time or multiple events. **Whether contemporary gene fusions tracked back to a fusion event in a common ancestor or whether they occurred independently was evaluated by phylogenetic analysis of the individual protein domains and by inspection of the inter-domain linker region. Linker regions were determined by multiple alignments of fusion sequences with corresponding free-standing domains present in the closest relatives to organisms that lack the gene fusions.**

Additional material

Additional file 1

Table S1, entitled "key to organism acronyms and sequence identifiers", is provided as supplementary material in an html document. This table contains a full collection of sequence data and annotations contained in this paper, and gi (gene identification) numbers are included and hyperlinked to facilitate access to the corresponding GenBank records. For future reference to a progressively updated table, refer to http://snp.lanl.gov/AroPath/SupplMaterials/TyrA/TableS1.html**.**

ACKNOWLEDGEMENTS

R. Jensen thanks the National Library of Medicine (Grant G13 LM008297) for partial support. This research is partially supported by the U. S. Army Research Institute of Infectious Diseases (USAMRIID) (Grant MIPR2MCTC32157).

REFERENCES

1. Xie G, Bonner CA, Jensen RA: **Cyclohexadienyl dehydrogenase from *Pseudomonas stutzeri* exemplifies a widespread type of tyrosine-pathway dehydrogenase in the TyrA protein family.** *Comp Biochem Physiol C Toxicol Pharmacol* 2000, **125**:65-83.
2. Jensen RA: **Tyrosine and phenylalanine biosynthesis: relationship between alternative pathways, regulation and subcellular location.** *Rec Adv Phytochem* 1986, **20**:57-82.
3. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a**

structural perspective. *J Mol Biol* 2001, **307**:1113-1143.

4. Teichmann SA, Rison SCG, Thornton JM, Riley M, Gough J, Clothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli***. *J Mol Biol* 2001, **311**:693-708.
5. Xie G, Keyhani N, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the dynamics of evolutionary change**. *Microbiol Mol Biol Rev* 2003, **67**:303-342.
6. Xie G, Bonner CA, Song J, Keyhani N, Jensen RA: **Inter-genomic displacement via lateral gene transfer of bacterial *trp* operons in an overall context of vertical genealogy**. *BMC Biology* 2004, **2**:15.
7. Gil R, Silva F, Zientz E, Delmotte F, Gonzalez-Candelas F, Latorre A, Rausell C, Kamerbeek J, Gadau J, Holldobler B, et al: **The genome sequence of *Blochmannia floridanus*: comparative analysis of reduced genomes**. *Proc Natl Acad Sci USA* 2003, **100**:9388-9393.
8. Blanc V, Gil P, Bamasjacques N, Lorenzon S, Zagorec M, Schleuniger J: **Identification and analysis of genes from *Streptomyces pristinaespiralis* encoding enzymes involved in the biosynthesis of the 4-dimethylamino-*L*-phenylalanine precursor of pristinamycin I**. *Mol Microbiol* 1997, **23**:191-202.
9. Xia T, Jensen RA: **A single cyclohexadienyl dehydrogenase specifies the prephenate dehydrogenase and arogenate dehydrogenase components of the dual pathways to *L*-tyrosine in *Pseudomonas aeruginosa***. *J Biol Chem* 1990, **265**:20033-20036.
10. Zhao G, Xia T, Ingram L, Jensen RA: **An allosterically insensitive class of cyclohexadienyl dehydrogenase from *Zymomonas mobilis***. *Eur J Biochem* 1993, **212**:157-165.
11. Jensen RA: **Enzyme recruitment in evolution of new function**. *Annu Rev Microbiol* 1976, **30**:409-425.
12. Subramaniam P, Bhatnagar R, Hooper A, Jensen RA: **The dynamic progression of evolved character states for aromatic amino acid biosynthesis in gram-negative bacteria**. *Microbiology* 1994, **140**:3431-3440.
13. Rippert P, Matringe M: **Molecular and biochemical characterization of an *Arabidopsis thaliana* arogenate dehydrogenase with two highly similar and active protein domains**. *Plant Mol Biol* 2002, **48**:361-368.
14. Rippert P, Matringe M: **Purification and kinetic analysis of the two recombinant arogenate dehydrogenase isoforms of *Arabidopsis thaliana***. *Eur J Biochem* 2002, **269**:4753-4761.
15. Xie G, Forst C, Bonner CA, Jensen RA: **Significance of two distinct types of tryptophan synthase beta chain in *Bacteria, Archaea* and higher plants**. *Genome Biol* 2001, **3**:Research0004.1-0004.13.
16. Champney WS, Jensen RA: **The enzymology of prephenate dehydrogenase in *Bacillus subtilis***. *J Biol Chem* 1970, **245**:3763-3770.
17. Bonner CA, Jensen RA, Gander JE, Kehani NO: **A core catalytic domain of the TyrA protein family: arogenate dehydrogenase from *Synechocystis***. *Biochem J* 2004.
18. Xie G, Bonner CA, Brettin TT, Gottardo R, Keyhani N, Jensen RA: **Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in *Xylella* species and heterocystous cyanobacteria**. *Genome Biol* 2003, **4**:R14.
19. Hall GC, Flick MB, Gherna RL, Jensen RA: **Biochemical diversity for biosynthesis of aromatic amino acids among the cyanobacteria**. *J Bacteriol* 1982, **149**:65-78.
20. Mavrodi DV, Ksenzenko VM, Bonsall RF, Cook RJ, Boronin AM, Thomashow LS: **A seven-gene locus for synthesis of phenazine-1-carboxylic acid by *Pseudomonas fluorescens* 2-79**. *J Bacteriol* 1998, **180**:2541-2548.
21. Pierson LS, Gaffney T, Lamb S, Gong F: **Molecular analysis of genes encoding phenazine biosynthesis in the biological control bacterium. *Pseudomonas aureofaciens* 30-84**. *FEMS Lett* 1995, **134**:299-307.
22. Eddy SR: **Profile-hidden Markov models**. *Bioinformatics* 1998, **14**:755-763.
23. Park J, Kaplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods**. *J Mol Biol* 1998, **284**:1201-1210.
24. Porat I, Waters BW, Teng Q, Whitman WB: **Two biosynthetic pathways for aromatic amino acids in the archaeon *Methanococcus maripaludis***. *J Bacteriol* 2004, **186**:4940-4950.
25. Calhoun DH, Bonner CA, Gu W, Xie G: **The emerging periplasm-localized subclass of AroQ**

chorismate mutases, exemplifed by those from *Salmonella typhimurium* and *Pseudomonas aeruginosa*. *Genome Biol* 2001, **2research0030.1-0030.16**.

26. Chen S, Vincent S, Wilson DB, Ganem B: **Mapping of chorismate mutase and prephenate dehydrogenase domains in the *Escherichia coli* T-protein**. *Eur J Biochem* 2003, **270**:757-763.
27. Ahmad S, Jensen RA: **The stable evolutionary fixation of a bifunctional tyrosine-pathway protein in enteric bacteria**. *Microbiol Lett* 1988, **52**:109-116.
28. Xie G, Brettin T, Bonner CA, Jensen RA: **Mixed-function supraoperons that exhibit overall conservation, albeit shuffled gene organization, across wide intergenomic distances within eubacteria**. *Microb Comp Genomics* 1999, **4**:5-28.
29. Xie G, Bonner CA, Jensen RA: **A probable mixed-function supraoperon in *Pseudomonas* exhibits gene organization features of both intergenomic conservation and gene shuffling**. *J Mol Evol* 1999, **49**:108-121.
30. Jensen RA, Ahmad S: **Nested gene fusions as markers of phylogenetic branchpoints in prokaryotes**. *Trends Ecol Evol* 1990, **5**:219-224.
31. Jensen RA, Gu W: **Evolutionary recruitment of biochemically specialized subdivisions of Family I within the protein superfamily of aminotransferases**. *J Bacteriol* 1996, **178**:2161-2171.
32. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches**. *J Mol Biol* 1999, **287**:1023-1040.
33. Henner D, Yanofsky C: ***Bacillus subtilis* and other gram-positive bacteria**. In: *Biochemistry, physiology, and molecular genetics* Edited by AL Sonenshein, J Hoch, R Losick. Washington, DC: ASM Press; 1993.
34. White RH: ***L*-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannashii***. *Biochemistry* 2004, **43**:7618-7627.
35. Ahmad S, Johnson JL, Jensen RA: **The recent evolutionary origin of the phenylalanine-sensitive isozyme of 3-deoxy-*D*-*arabino*-heptulosonate 7-phosphate synthase in the enteric lineage of bacteria**. *J Mol Evol* 1987, **25**:159-167.
36. Pittard AJ: **Biosynthesis of the amino acids**. In: *Escherichia coli and Salmonella: cellular and molecular biology.* Edited by F Neidhardt, Curtiss III, R, Ingraham, JL, Lin, e. c. C. , Low, KB, Magasanik, B, Remikoff, w. s., Riley, M., Schaechter, M., Umbarger, H. E. pp. 458-484: American Society for Microbiology; 1996: 458-484.
37. Katayama T, Suzuki H, Koyanagi T, Kumagai H: **Cloning and random mutagenesis of the *erwinia herbicola tyrR* gene for high-level expression of tyrosine phenol-lyase**. *Appl Envir Microbiol* 2000, **66**:4764-4771.
38. Bai Q, Somerville R: **Integration host factor and cyclic AMP receptor proein are required for TyrR-mediated activation of *tpl* in *Citrobacter freundii***. *J Bacteriol* 1998, **180**:6173-6186.
39. Zhao S, Somerville RL: **Isolated operator binding and ligand response domains of the TyrR protein of *Haemophilus influenzae* associate to reconstitute functional repressor**. *J Biol Chem* 1999, **274**:1842-1847.
40. Arias-Barrau E, Olivera E, Luengo J, Fernandez C, Galan B, Garcia J, Diaz E, Miñambres B: **The homogentisate pathway: a central catabolic pathway involved in the degradation of *L*-phenylalanine, *L*-tyrosine, and 3-hydroxyphenylacetate in *Pseudomonas putida***. *J Bacteriol* 2004, **186**:5062-5077.
41. Dähnhardt D, Falk J, Appel J, van der Kooij A, Schulz-Friedrich R, Krupinska K: **The hydroxyphenylpyruvate dioxygenase from *Synechocystis* sp. PCC 6803 is not required for plastoquinone biosynthesis**. *FEBS Lett* 2002, **523**:177-181.
42. Song J, Jensen RA: **PhhR, a divergently transcribed activator of the phenylalanine hydroxylase gene cluster of *Pseudomonas aeruginosa***. *Mol Microbiol* 1996, **22**:497-507.
43. Zhao G, Xia T, Song J, Jensen R: ***Pseudomonas aeruginosa* possesses homologues of mammalian phenylalanine hydroxylase and 4a-carbinolamine dehydratase/DCoH as part of a three-component gene cluster**. *Proc Natl Acad Sci USA* 1994, **91**:1366-1370.
44. Tropel D, van der Meer J: **Bacterial transcriptional regulators for degradation pathways of aromatic compounds**. *Microbiol Mol Biol Rev* 2004, **68**:474-500.
45. Chaney M, Grande R, Wigneshweraraj S, Cannon W, Casaz P, Gallegos M-T: **Binding of**

transcriptional activators to sigma 54 in the presence of the transition state analog ADP-aluminum fluoride: insights into activator mechanochemical action. *Genes Dev* 2001, **15**:2282-2294.

46. Yanofsky C: **The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis*.** *Trends Genet* 2004, **20**:367-374.
47. Chenna R, Sugawara H, Koike T, Lopez R, Gibson T, Higgins D, Thompson J: **Multiple sequence alignment with the Clustal series of programs.** *Nucl Ac Res* 2003, **31**:3497-3500.
48. Felsenstein J: **PHYLIP-Phylogeny Inference Package (version 3.2).** *Cladistics* 1989, **5**:164-166.
49. Cai W, Pei J, Grishin NV: **Reconstruction of ancestral protein sequences and its applications.** *BMC Evol Biol* 2004, **4**:33.
50. Eddy S: **HMMER package (http://hmmer.wustl.edu).** 1995.
51. Lingens F, Göbel W, Üsseler H: **Regulation der biosynthesis der aromatischen aminosauren in *Saccharomyces cerevisiae*, I. Hemmung der Enzymaktivitaten (Feedback-Wirkung).** *Biochem Z* 1966, **346**:357-67.
52. Fazel A, Jensen R: **Obligatory biosynthesis of *L*-tyrosine via the pretyrosine branchlet in coryneform bacteria.** *J Bacteriol* 1979, **138**:805-815.
53. Fazel AM, Bowen JR, Jensen RA: **Arogenate (pretyrosine) is an obligatory intermediate of *L*-tyrosine biosynthesis: confirmation in a microbial mutant.** *Proc Natl Acad Sci USA* 1980, **77**:1270-1273.

FIGURE LEGENDS

Figure 1

Composite of alternative biochemical routes from chorismate (CHA) to *L*-tyrosine (TYR) in nature. An antibiotic synthesis branch from CHA is also shown (dimmed). PPA may be transaminated by prephenate aminotransferase (PAT) to yield *L*-arogenate (AGN). The four TyrA homologs and the reactions they catalyze are colored differently. Arogenate dehydrogenase ($TyrA_a$) converts AGN to TYR. Alternatively, prephenate dehydrogenase ($TyrA_p$) converts PPA to 4-hydroxyphenylpyruvate (HPP) which is then transaminated to TYR via an homolog of TyrB, AspC, HisH, or Tat [31]. A broad-specificity cyclohexadienyl dehydrogenase ($TyrA_c$) is competent to catalyze either the $TyrA_a$ or the $TyrA_p$ reaction. PapC converts the 4-amino analog of PPA to the 4-amino analog of HPP. AroQ, AroH, and AroR are distinct homologs known to exist in nature for performance of the chorismate mutase reaction. Other abbreviations: AA, amino acid donor, KA, keto-acid accepter.

Figure 2

Phylogenetic tree for the TyrA Superfamily. Acronyms used for the various organisms are given in alphabetical order in Table 2. (A more extensive key to organism acronyms that lists the substrate specificities of their TyrA proteins and which is hyperlinked to the individual GenBank records is given in Table S1). The four experimentally established $TyrA_a$ lineages are highlighted in **blue**. A single large clade of Gram-positive bacterial species represents the sole established $TyrA_p$ specificity on the protein tree. Although *Sce* $TyrA_x$ has been characterized as a $TyrA_p$ protein, this was reported [51] long before it was recognized that prephenate preparations were often contaminated with AGN. All other proteins are $TyrA_c$ or $TyrA_{c\_?}$, except for PapC proteins from *Sco*, *Plu*, and *Spr* (see

text). All proteins having an aspartate residue homologous to D-32 of the *E. coli* $TyrA_{c\_?\_NAD}$ domain are presumed specific for NAD^+, and these are indicated with orange highlighting. TyrA proteins that are not highlighted are either $NADP^+$-specific, broadly specific for either pyridine nucleotide, or the exact cofactor specificity is unknown (see Table 1). Fusion of TyrA domains with other catalytic domains is indicated within grey boxes (AroQ•TyrA, TyrA•AroF, $HisH_b$•TyrA, and TyrA•AroQ•PheA•ACT) using the convention of a bullet to represent the interdomain area. The boxes overlap the relevant lineages. TyrA proteins having carboxy-terminal fusions with regulatory domains (TyrA•ACT and TyrA•REG) are also shown. Filled circles at node positions denote bootstrap values of 998-to-1000. The distance scale bar represents substitutions per site.

Figure 3

Multiple alignment of the HMM consensus sequences obtained for different substrate specificity groupings. Invariant anchor residues are highlighted in yellow, conserved residues in grey. The aspartate residues associated with specificity for NAD^+ are shown in **red**.

FIGURE 4

Context of gene organization for *tyrA*, mapped on the 16S rRNA tree of the domain *Bacteria*. *pheA*, $hisH_b$, *tyrA*, and *aroF* are color-coded. Lineages typified by complete dispersal of aromatic-pathway genes are indicated with divergent dashed lines. Consensus gene organizations are shown for alpha, beta, and gamma divisions of the Proteobacteria. The gamma division is further subdivided to yield the "upper-gamma" organisms and the "lower-gamma" (enteric lineage) organisms.

Figure 5

Zoom-in from Fig. 4 showing *tyrA* synteny for the beta-proteobacteria and the "upper" gamma proteobacteria. The tree shown, based upon 16S rRNA sequences of the indicated organisms, indicates correct branching orders but is not strictly correct in proportion. Circled numbers at the top indicate deduced evolutionary events for the beta-proteobacteria (see top of Table 4), whereas circled numbers at the bottom (see bottom of Table 4) correspond to deduced evolutionary events for the "upper" gamma proteobacteria. Gene organizations of organisms indicated are shown on the right. See Table 2 for organismal acronyms.

Figure 6

Zoom-in from Fig. 4 showing *tyrA* synteny for the "lower" gamma-proteobacteria (enteric lineage). Deduced phylogenetic events numbered on the left are described in Table 5. The branching position for *Buchnera* is as suggested in ref. [5].

Figure 7

Distribution of modules of aromatic catabolism mapped on a 16S rRNA tree. The Phh module (orange) consists of phenylalanine hydroxylase (PhhA) carbinolamine dehydratase (PhhB), and tyrosine aminotransferase (not shown, see Text), and accomplishes the overall conversion of PHE to 4-hydroxyphenylpyruvate. The Hpd module (yellow) is 4-hydroxyphenylpyruvate dioxygenase, which converts 4-hydroxyphenylpyruvate to homogentisate. The Hmg module (**blue**) catalyzes the 3-step conversion of homogentisate to acetoacetate and fumarate. The distribution of PhhR and TyrR is shown in boxes. Homologs of uncertain function are depicted by

boxed question marks. In some cases the HmgC member is shaded light blue to indicate that the gene encoding this isomerase could not be found. This is probably encoded by an as yet unknown analog. Branches that end on the right with double bars or triple bars are longer by distances equal to 25% or 12% of the scale bar.

Figure 8

Protein tree of close TyrR homologs. Nodes supported by bootstrap values of 998 or more are marked with solid circles, and the bootstrap values for nodes internal to these are shown. Generic names relevant to the organism abbreviations can be viewed in Fig. 7. A conserved region containing the σ^{54} contact motif **GAFTGA** is highlighted as a yellow band. Imperfect residues in this region are shown in lower-case fonts. TyrR and PhhR are regulators of σ^{70} and σ^{54} promoters, respectively. Residue numbers are shown at the right. Four σ^{54} proteins of unknown function have very long branches, and the gaps in branch continuity that are drawn represent a scale-bar distance of 0.1.

Table 1. Abbreviations used to designate substrate specificities of *tyrA*/TyrA homologs

Abbreviation[a]		Description of specificity[b]
Gene	Gene Product	
tyrA$_x$	TyrA$_x$	Specificity for cyclohexadienyl substrate is unknown
tyrA$_c$	TyrA$_c$	Broad-specificity cyclohexadienyl dehydrogenase (CDH)
tyrA$_p$	TyrA$_p$	Narrow-specificity prephenate dehydrogenase (PDH)
TyrA$_{c_?}$	TyrA$_{c_?}$	Broad-specificity cyclohexadienyl dehydrogenase having catalytic-core indels in correlation with an extra-core extension
tyrA$_a$	TyrA$_a$	Narrow-specificity arogenate dehydrogenase (ADH)

$_{NAD}$*tyrA*$_a$	$_{NAD}$TyrA$_a$	TyrA homolog is AGN-specific and NAD^+-specific
$_{NADP}$*tyrA*$_a$	$_{NADP}$TyrA$_a$	TyrA homolog is AGN-specific and $NADP^+$-specific
$_{NAD(P)}$ *tyrA*$_a$	$_{NAD(P)}$TyrA$_a$	TyrA homolog is AGN-specific but utilizes either NAD^+ or $NADP^+$
$_x$*tyrA*$_x$	$_x$TyrA$_x$	Specificity for both the cyclohexadienyl and pyridine nucleotide substrates is unknown

[a]Abbreviations in the upper-table box indicate the specificities for the cyclohexadienyl substrate. Abbreviations in the lower-table box indicate specificities for both cyclohexadienyl (right subscripts) and pyridine nucleotide substrates (left subscripts). Combinations not shown can be deduced from the examples given, e.g., a TyrA homolog specific for prephenate and NAD would be designated $_{NAD}$TyrA$_p$.

[b]The abbreviations CDH, PDH, and ADH (shown parenthetically) have been used frequently in the literature.

TABLE 2. KEY TO ORGANISM ACRONYMS

Organism	Abbreviation
Actinobacillus actinomycetemcomitans	Aac
Acinetobacter sp.	Acs
Alcaligenes bronchisepticus	Abr
Anabaena sp.	Asp
Anabaena sp.	Asp_2
Arabidopsis thaliana	Ath
Archaeoglobus fulgidus	Afu
Agrobacterium tumefaciens	Atu
Azotobacter vinelandii	Avi
Bacillis anthracis	Ban
Bacillus cereus	Bce
Bacillus halodurans	Bha
Bacillus stearothermophilus	Bst
Bacillis subtilis	Bsu
Bacillus thuringiensis israelensis	Bth
Bifidobacterium longum	Blo
Blochmannia floridanus	Bfl
Burkholderia cepacia	Buc
Burkholderia pseudomallei	Bps
Campylobacter jejuni	Cje
Corynebacterium glutamicum	Cgl
Enterococcus faecalis	Efa
Enterococcus faecium	Enf
Erwinia herbicola	Ehe
Escherichia coli	Eco
Haemophilus influenzae	Hin
Helicobacter pylori	Hpy
Klebsiella pneumoniae	Kpn
Lycopersicon esculentum	Les
Listeria innocua	Lin
Listeria monocytogenes	Lmo
Methanobacterium thermoautotrophicum	Mth
Methanococcus jannaschii	Mja
Methanopyrus kandleri AV19	Mka
Methanosarcina barkeri	Mba
Microbulbifer degradans	Mde
Mycobacterium tuberculosis	Mtu
Neisseria gonorrhoeae	Ngo
Nitrosomonas europaea	Neu
Nostoc punctiforme	Npu
Nostoc punctiforme	Npu_2
Sphingomonas aromaticivorans	Nar
Oryza sativa ssp. Japonica	Osa

Pantoea agglomerans	Pag
Pasteurella multocida	Pmu
Prochlorococcus marinus CCMP1378 MED4	Pma_C
Prochlorococcus marinus MIT9313	Pma_M
Pseudomonas aeruginosa	Pae
Pseudomonas fluorescens	Pfl
Pseudomonas putida	Ppu
Pseudomonas stutzeri	Pst
Ralstonia eutropha	Reu
Ralstonia solanacearum	Rso
Rhodobacter capsulatus	Rca
Rhodobacter sphaeroides	Rsp
Rhodopseudomonas palustris	Rpa
Rhodospirillum rubrum	Rru
Saccharomyces cerevisiae	Sce
Salmonella typhimurium	Sty
Schizosaccharomyces pombe	Spo
Shewanella putrefaciens	Spu
Staphylococcus aureus	Sau
Streptococcus gordonii	Sgo
Streptococcus pneumoniae	Spn
Streptomyces coelicolor	Sco
Streptomyces coelicolor	Sco_2
Streptomyces pristinaespiralis	Spr
Synechococcus sp. WH8102	Scc_W
Synechococcus sp. PCC7002	Scc_7
Synechocystis sp. PCC6803	Ssp
Vibrio cholerae	Vch
Xanthomonas campestris	Xca
Xylella fastidiosa (almond strain)	Xal
Yersinia enterocolitica	Yen
Zymomonas mobilis	Zmo

Table 3. Cyclohexadienyl substrates and inhibitors of TyrA proteins possess identical sidechains

Organism	Co-Factor	Substrate(s)	Inhibitor(s)[a]	Reference
Synechocystis sp.	$NADP^+$	AGN	TYR	[17]
Arabidopsis thaliana	$NADP^+$	AGN	TYR	[13, 14]
Nitrosomonas europaea	$NADP^+$	AGN	None	[12]
Corynebacterium glutamicum	$NAD(P)^+$	AGN	None	[52, 53]
Neisseria gonorrhoeae	NAD^+	PPA[b]	HPP	[12]
Pseudomonas stutzeri	NAD^+	PPA/AGN	HPP/TYR	[1]
Pseudomonas aeruginosa	NAD^+	PPA/AGN	HPP/TYR	[9]
Zymomonas mobilis	NAD^+	PPA/AGN	None	[10]

[a]Abbreviation: HPP, 4-hydroxyphenylpyruvate. [b]This $TyrA_c$ enzyme has an overwhelming preference for PPA, but will use AGN poorly.

Table 4. Key to evolutionary events asserted in Figure 5

Number		Evolutionary event(s) proposed
Beta	1	Dispersal of the four aromatic-pathway genes away from one another and away from *gyrA* > *serC* and from *cmk* > *rspA* > *himD*; inversion of *aroF* with respect to *cmk*.
		Complete dispersal of all nine genes originally in the *gyrA*/*himD* linkage group.
	2	Insertion of *serA* after *serC*[a]; separation of *tyrA* and *aroF* to yield a 6-gene unit and a 4-gene unit.
	3	Expulsion of $hisH_b$; insertion of 'ORF' after *serC*.
		Fusion of *tyrA* with *aroF*.
	4	Loss of $hisH_b$ from genome.
	5	
	6	
Upper-Gamma	1	Insertion of *serA* after *serC*[a]; insertion of $aroA_{Ia}$ after $hisH_b$.
	2	Translocation of $hisH_b$ and *tyrA* to other regions, leaving two separated 3-gene units.
	3	Fusion of *tyrA* with *aroF*.
	4	Loss of $hisH_b$.
	5	N-terminal deletion of •*aroF* domain, and acquisition of new *aroF* gene (probable LGT).
		Separation of *cmk* > *rpsA* > *himD* from *aroQ*•*pheA* > *tyrA*•*aroF*.
	6	Insertion of 4 unknown genes between *gyrA* and *serC* in opposite orientation and separation of *gyrA* > ORF > ORF > ORF > *serC* from *aroQ*•*pheA* > *tyrA*•*aroF*.
	7	Loss of *himD*; translocation of *serC* away from *gyrA* and *aroQ*•*pheA*.
	8	

[a]Since both *Nitrosomonas* (beta-proteobacteria) and *Acidothiobacillus* (upper-gamma proteobacteria) emerge at deep positions in the tree of Fig. 4, an almost equally parsimonius possibility is that the ancestral *serA* was retained in this syntenic position in these two genera, but was transposed elsewhere shortly after early divergence.

Table 5. Key to evolutionary events asserted in Figure 6

Number	Evolutionary events proposed
1	Escape of *aroQ•pheA* and *tyrA* from the ancestral *gyrA* > *serC* > *aroQ•pheA* > *hisH*$_b$ > *tyrA* > *aroF* > *cmk* > *rpsA* > *himD* supraoperon. Origin of *aroQ•tyrA* fusion. Origin of *aroA*$_{Ia\_Y}$ > *aroQ•tyrA* operon. Addition of *tyrR*. Addition of third *aroA*$_{Ia}$ species: *aroA*$_{Ia\_F}$.
2	Fusion of *aroQ•pheA* with *aroA*$_{I\beta}$ pseudogene of unknown origin. Replacement of *hisH*$_b$ by *aspC* duplicate linked with three ORFs.
3	Dissociation of *gyrA* and *serC*.
4	Removal of all genes intervening between *aroQ•pheA* and *aroQ•tyrA*.
5	Dissociation of *aroF* from both *serC* and *cmk* > *rpsA* > *himD*. Insertion of *trpR* and ____ within the intervening region between *aroQ•pheA* and *aroQ•tyrA*.
6	Dissociation of *serC* > *hisH*$_b$ > *aroF* from *cmk* > *rpsA* > *himD*.
7	Loss of *aroA*$_{Ia\_Y}$ from *tyr* operon.
8	*aroF* becomes dissociated from *hisH*$_b$, and *aroA*$_{Ia\_Y}$ is removed from the *tyrA* operon.
9	*ORF* > *gyrA* is inserted after *aroF*.
10	*aroQ•tyrA* becomes a pseudogene.
11	*hisH*$_b$ is lost.
12	*himD* is lost.
13	*cmk*, *himD* and *aroA*$_{Ia\_Y}$ > *aroQ•tyrA* are lost.
14	*aroF*, *himD*, *aroQ•pheA*, and *aroA*$_{Ia\_Y}$ > *aroQ•tyrA* are lost.
15	All intervening genes between *aroQ•pheA* and *aroQ•tyrA* are eliminated.
16	*pheA* domain of *aroQ•pheA* becomes a pseudogene.
17	Insertion of *ycaL* between *aroF* and *cmk*.
18	Insertion of ORF between *aroF* and *ycaL*.
19	Insertion of ORF between *aroQ•pheA* and *aroQ•tyrA*.

Table 6. Putative attenuators[a] associated with *tyrA*

Organism	Gene organization[b]	Fig[c]
Agrobacterium tumefaciens	⌐ *hisH$_b$* > ***tyrA***	4[d]
Bacillus anthracis	⌐ *aroG* > *hisH $_b$*> ***tyrA*** > *aroF*	(11)
Bacillus cereus	⌐ *aroG* > *hisH$_b$* > ***tyrA*** > *aroF*	(11)
Bacillus halodurans	⌐ ***tyrA*** > *aroF*	(11)
Bacteroides thetaiotaomicron	⌐ *pheA* > *hisH$_b$* > *aroA$_{1\beta}$*•aroQ > ***tyrA***	4
Bordetella parapertussis	⌐ *gyrA* > *serC* > *aroQ*•*pheA* > ***tyrA*** > *aroF* > *cm k*> *rpsA* > *himD*	5
Desulfovibrio vulgaris	⌐ *aroA'* > *aroB'* > *aroQ*•*pheA* > *aroF* > ***tyrA*** > [*trp* operon]	4
Enterococcus faecalis	⌐ *aroD* > *aroA$_{1\beta}$*> *aroB* > *aroG* > ***tyrA*** > *aroF* > *aroE* > *pheA*	4
Lactococcus lactis	⌐ *ysaA* > *blrG* > *kinG* > ***tyrA*** > *aroF* > *aroE* > *pheA*	4
Lactobacillus plantarum	⌐ ORF > *aroG* > *ORF* > *aroF* > ***tyrA*** > *aroE*	
Listeria innocua	⌐ *aroG* > *aroB* > *aroH* > *hisH $_b$*> ***tyrA*** > *aroF*	4,(11)
Streptococcus pneumoniae	⌐ *ORF*>*aroC$_I$*>*aroD*>*aroB*>*aroG*>***tyrA***> ⌐ *ORF*>*aroF*>*aroE*>*pheA*	4
Thermoanaerobacter tencongensis	⌐ *pheA* > *aroA$_{1\beta}$*> ***tyrA*** > *aroF* > *ORF* > *ORF*	
Thermus thermophilus	⌐ *aroA$_{1\beta}$* > ***tyrA***	
Vibrio parahaemolyticus	⌐ *aroA$_{1a_Y}$* > ***tyrA***	6
Vibrio vulnificus	⌐ *aroA$_{1a_Y}$* > ***tyrA***	6

[a]Attenuators were extracted from the website of Merino and Yanofsky (http://cmgm.stanford.edu/~merino/). These are included within a larger list of attenuators relevant to general aromatic biosynthesis that were extracted from the Merino and Yanofsky database and placed on our website (http://snp.lanl.gov/AroPath/SupplMaterials/TyrA/Table6.html). Links are provided for viewing the complete data, including a visualization of the putative attenuator structures.

[b]Symbol used for attenuator: ⌐ *aroA'* and *aroB'* refer to alternative biochemical steps recently reported for formation of dehydroquinate from aspartate semialdehyde and ketohexose 1-phosphate [34]

[c]Refers to figures within this manuscript or if enclosed within parentheses, to the figure in ref. [5].

[d]See the consensus gene organization for a-proteobacteria.

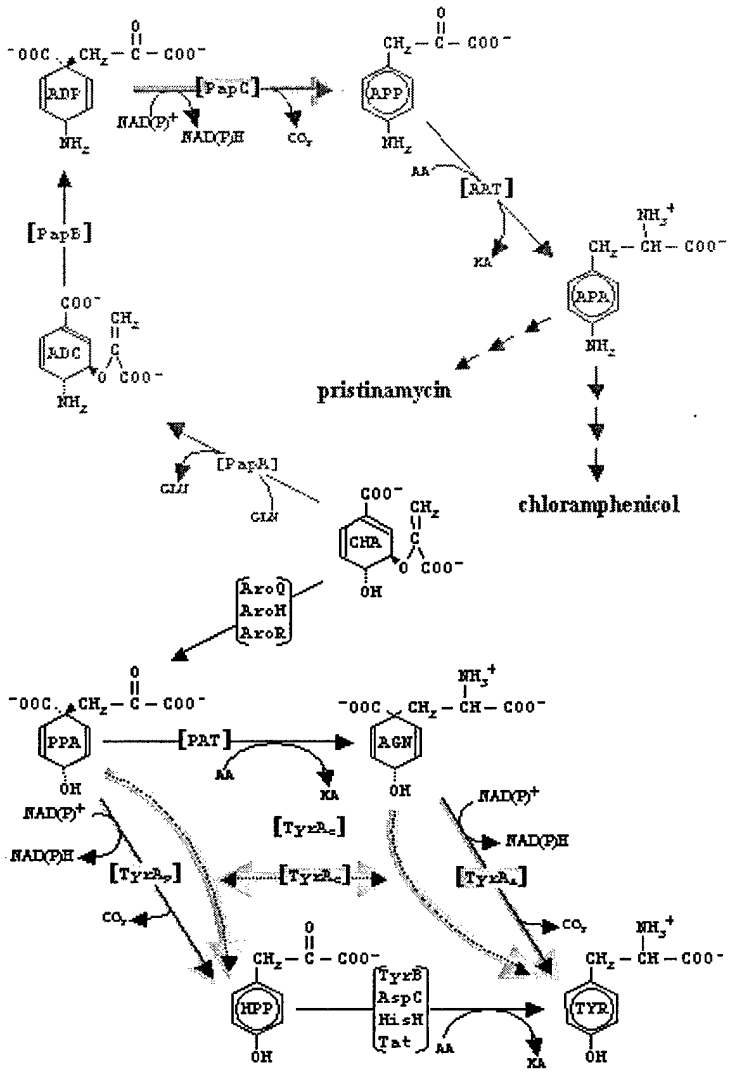

ADF
[PapC]
NAD(P)+
NAD(P)H
CO2
APP
AA
[APT]
KA
APA
pristinamycin
[PapB]
ADC
[PapA]
GLU
GLN
CHA
chloramphenicol
AroQ
AroH
AroE
PPA
[PAT]
AGN
[TyrA]
HPP
TyrB
AspC
HisH
Tat
TYR

Figure 1

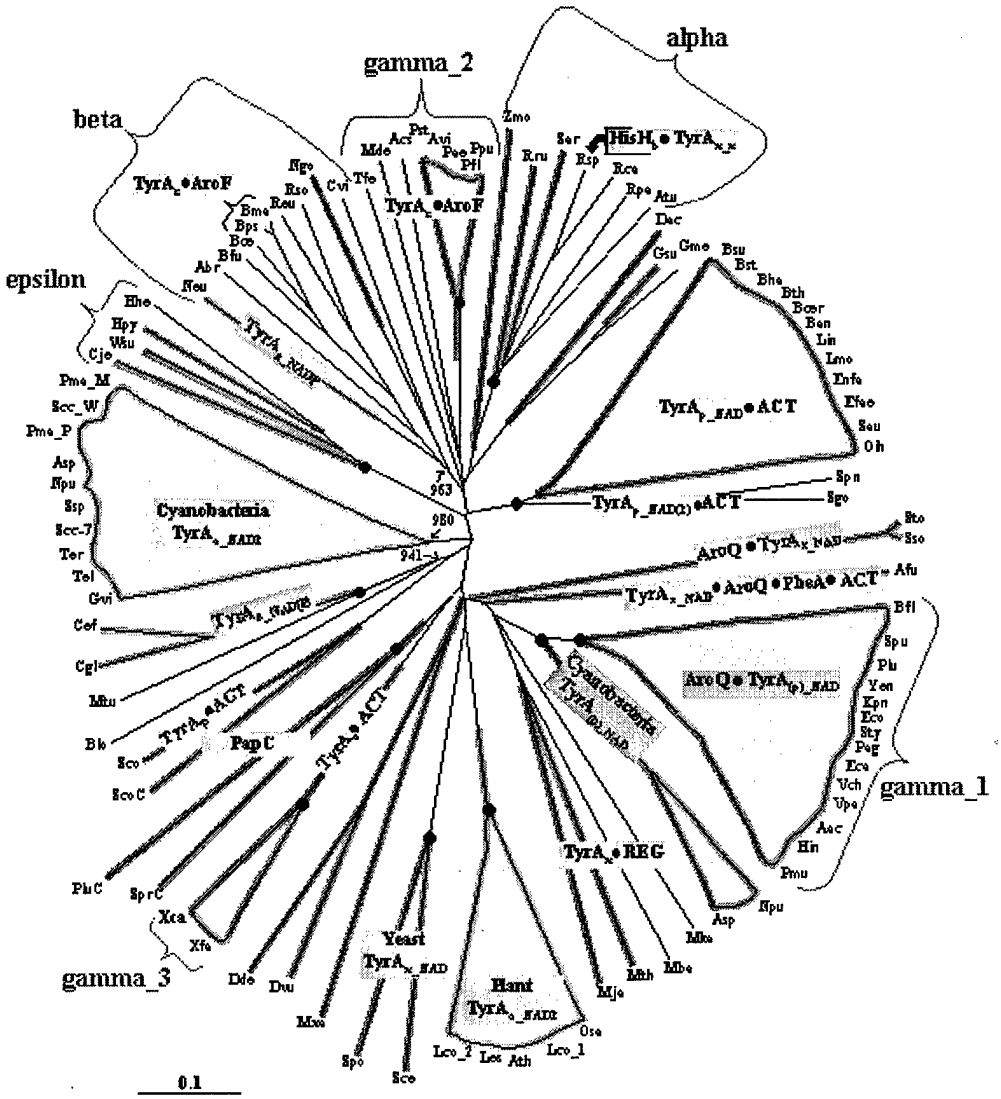

alpha
gamma_2
beta
epsilon
gamma_1
gamma_3
Cyanobacteria
Yeast
Plant
PapC
0.1

Figure 2

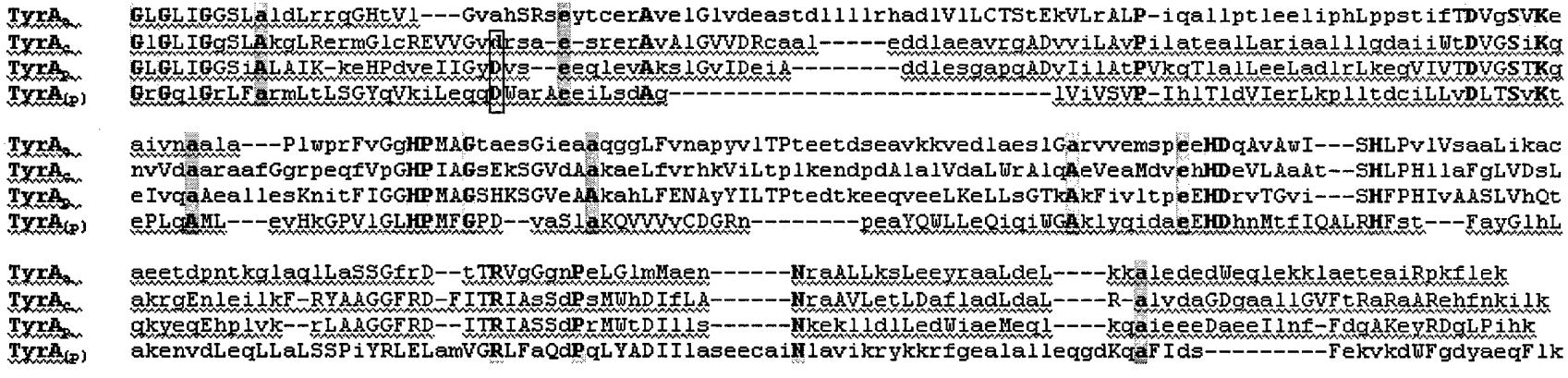

TyrA
TyrA
TyrA
TyrA(p)
TyrA
TyrA
TyrA
TyrA(p)
TyrA
TyrA
TyrA
TyrA(p)

Figure 3

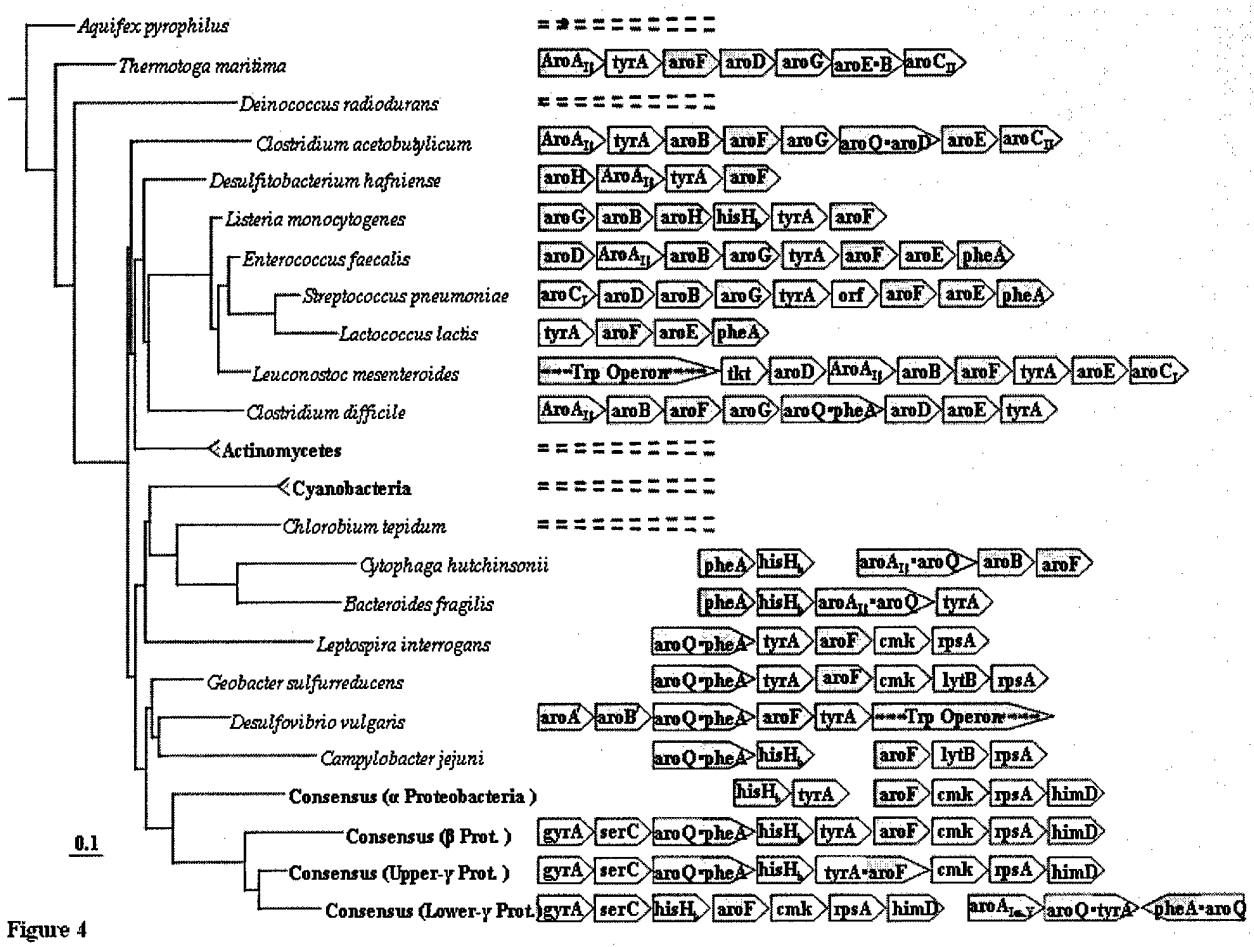

Aquifex pyrophilus
Thermotoga maritima
Deinococcus radiodurans
Clostridium acetobutylicum
Desulfitobacterium hafniense
Listeria monocytogenes
Enterococcus faecalis
Streptococcus pneumoniae
Lactococcus lactis
Leuconostoc mesenteroides
Clostridium difficile
Actinomycetes
Cyanobacteria
Chlorobium tepidum
Cytophaga hutchinsonii
Bacteroides fragilis
Leptospira interrogans
Geobacter sulfurreducens
Desulfovibrio vulgaris
Campylobacter jejuni
Consensus (α Proteobacteria)
Consensus (β Prot.)
Consensus (Upper-γ Prot.)
Consensus (Lower-γ Prot.)
0.1
aroA
tyrA
aroF
aroD
aroG
aroE-B
aroC
aroB
aroH
hisH
orf
pheA
tkt
aroQ-pheA
aroQ-tyrA
tyrA-aroF
aroA-aroQ
cmk
lytB
rpsA
himD
gyrA
serC
Trp Operon

Figure 4

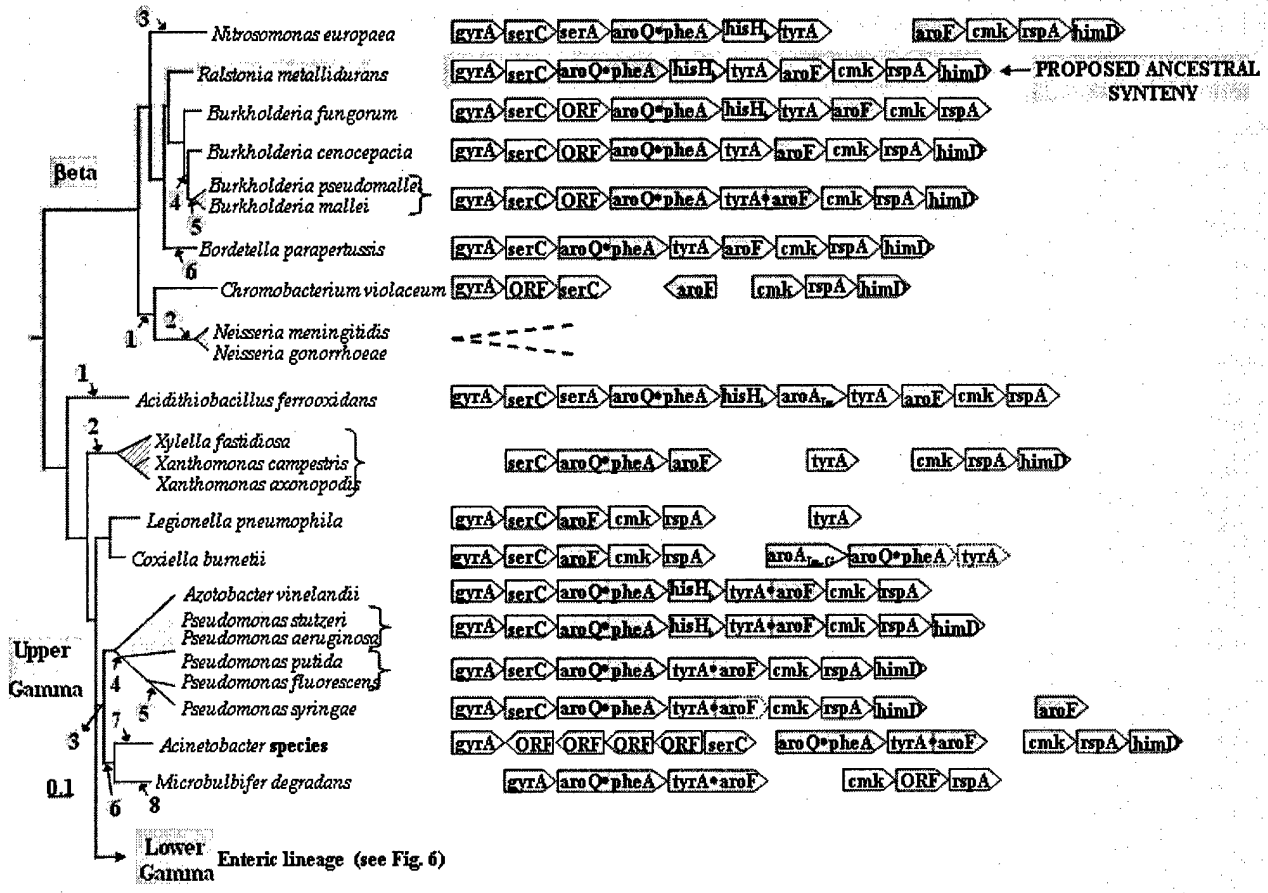

Nitrosomonas europaea
Ralstonia metallidurans
Burkholderia fungorum
Burkholderia cenocepacia
Burkholderia pseudomallei
Burkholderia mallei
Bordetella parapertussis
Chromobacterium violaceum
Neisseria meningitidis
Neisseria gonorrhoeae
Acidithiobacillus ferrooxidans
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas axonopodis
Legionella pneumophila
Coxiella burnetii
Azotobacter vinelandii
Pseudomonas stutzeri
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas fluorescens
Pseudomonas syringae
Acinetobacter species
Microbulbifer degradans
PROPOSED ANCESTRAL SYNTENY
Beta
Upper Gamma
Lower Gamma
Enteric lineage (see Fig. 6)
0.1

Figure 5

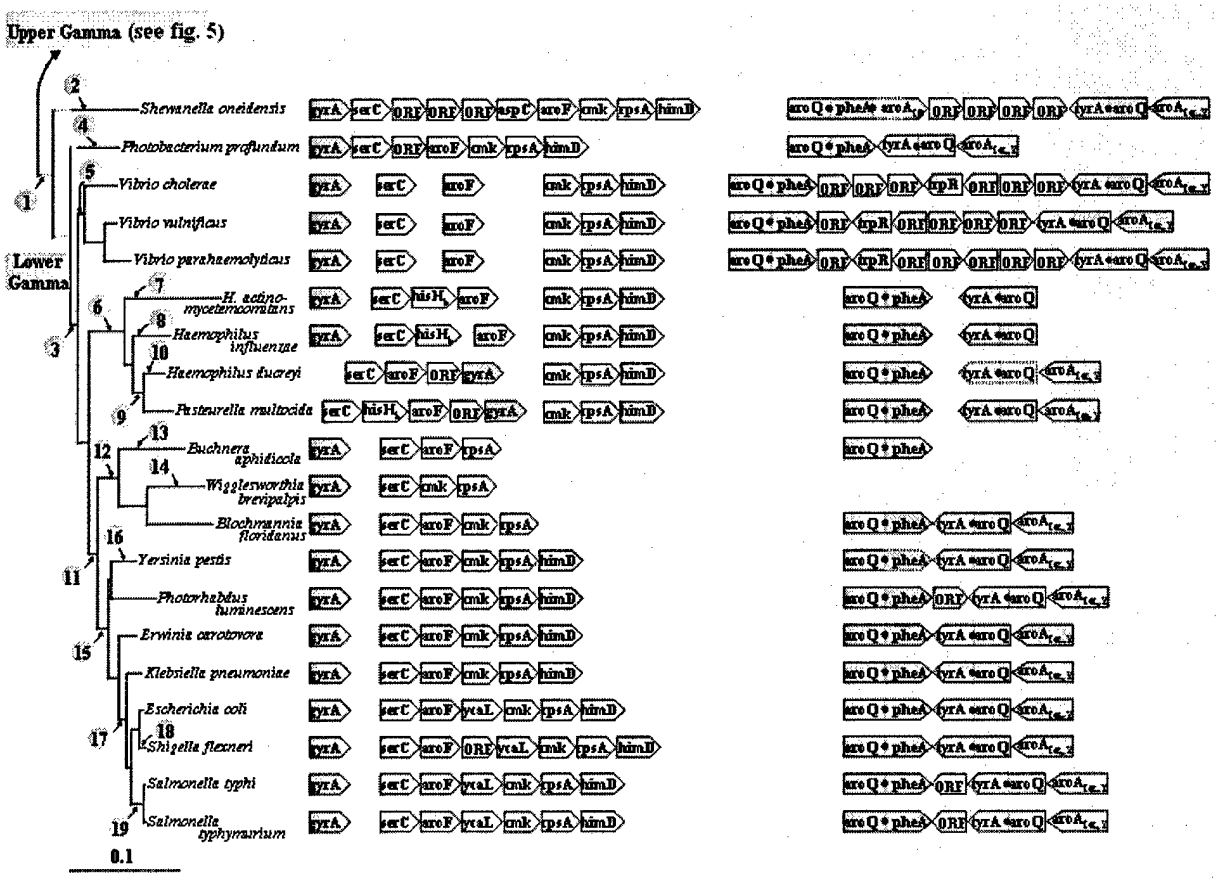

Upper Gamma (see fig. 5)
Lower Gamma
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio vulnificus
Vibrio parahaemolyticus
H. actinomycetemcomitans
Haemophilus influenzae
Haemophilus ducreyi
Pasteurella multocida
Buchnera aphidicola
Wigglesworthia brevipalpis
Blochmannia floridanus
Yersinia pestis
Photorhabdus luminescens
Erwinia carotovora
Klebsiella pneumoniae
Escherichia coli
Shigella flexneri
Salmonella typhi
Salmonella typhymurium
0.1

Figure 6

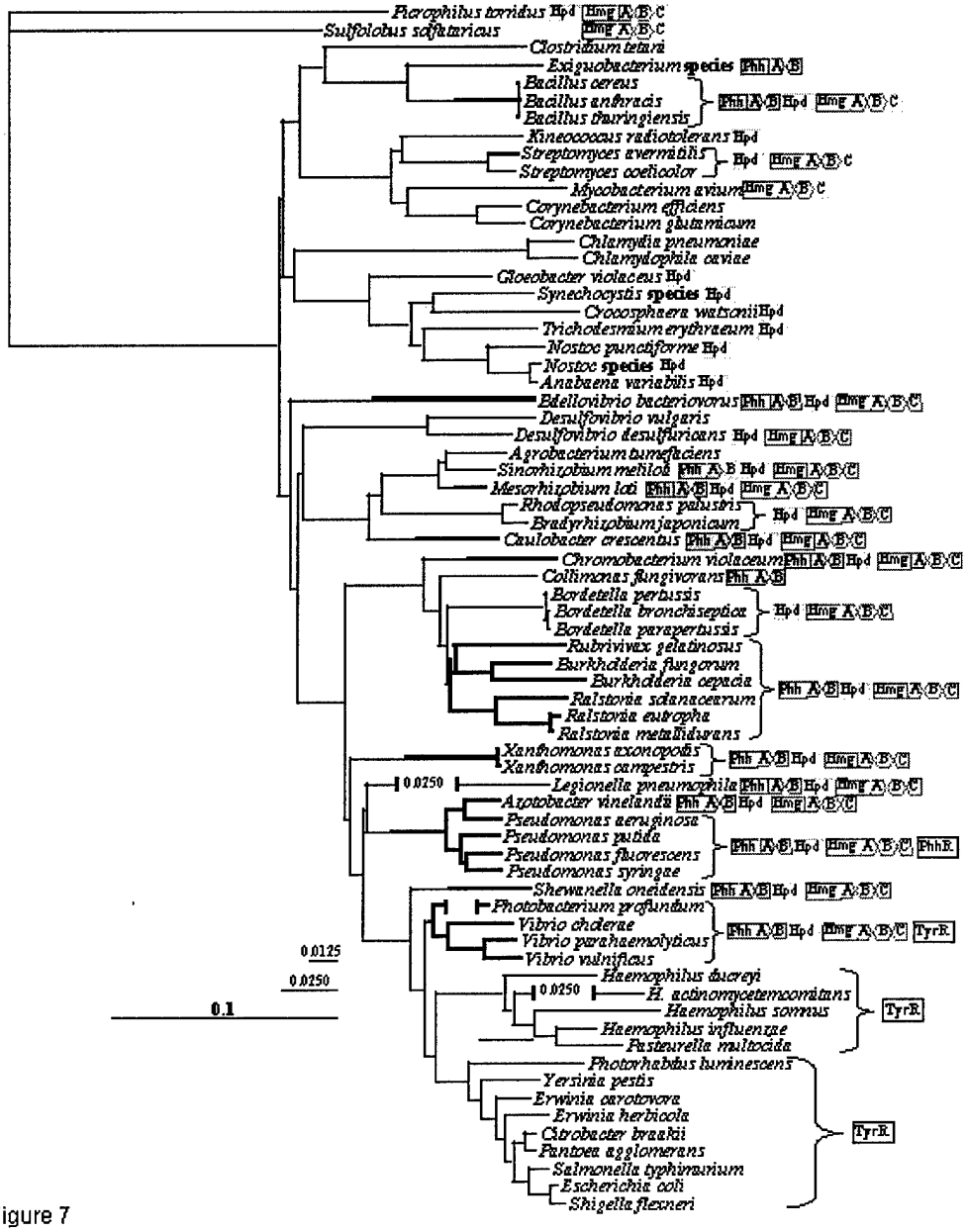
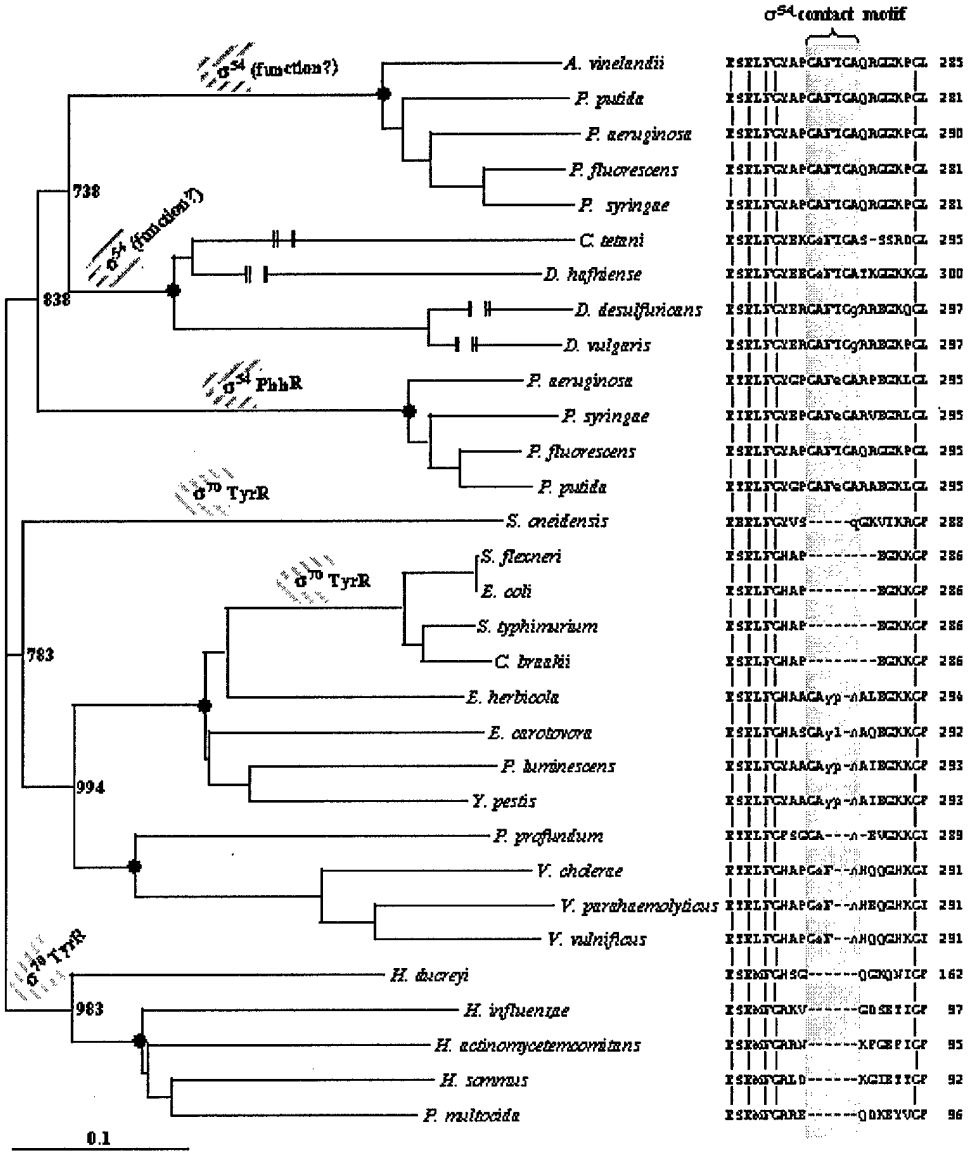

Picrophilus torridus Hpd HmgA/B/C
Sulfolobus solfataricus HmgA/B/C
Clostridium tetani
Exiguobacterium species PhhA/B
Bacillus cereus
Bacillus anthracis
Bacillus thuringiensis PhhA/B Hpd HmgA/B/C
Kineococcus radiotolerans Hpd
Streptomyces avermitilis
Streptomyces coelicolor Hpd HmgA/B/C
Mycobacterium avium HmgA/B/C
Corynebacterium efficiens
Corynebacterium glutamicum
Chlamydia pneumoniae
Chlamydophila caviae
Gloeobacter violaceus Hpd
Synechocystis species Hpd
Crocosphaera watsonii Hpd
Trichodesmium erythraeum Hpd
Nostoc punctiforme Hpd
Nostoc species Hpd
Anabaena variabilis Hpd
Bdellovibrio bacteriovorus PhhA/B Hpd HmgA/B/C
Desulfovibrio vulgaris
Desulfovibrio desulfuricans Hpd HmgA/B/C
Agrobacterium tumefaciens
Sinorhizobium meliloti PhhA/B Hpd HmgA/B/C
Mesorhizobium loti PhhA/B Hpd HmgA/B/C
Rhodopseudomonas palustris
Bradyrhizobium japonicum Hpd HmgA/B/C
Caulobacter crescentus PhhA/B Hpd HmgA/B/C
Chromobacterium violaceum PhhA/B Hpd HmgA/B/C
Collimonas fungivorans PhhA/B
Bordetella pertussis
Bordetella bronchiseptica
Bordetella parapertussis Hpd HmgA/B/C
Rubrivivax gelatinosus
Burkholderia fungorum
Burkholderia cepacia
Ralstonia solanacearum
Ralstonia eutropha
Ralstonia metallidurans PhhA/B Hpd HmgA/B/C
Xanthomonas axonopodis
Xanthomonas campestris PhhA/B Hpd HmgA/B/C
0.0250
Legionella pneumophila PhhA/B Hpd HmgA/B/C
Azotobacter vinelandii PhhA/B Hpd HmgA/B/C
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas fluorescens
Pseudomonas syringae PhhA/B Hpd HmgA/B/C PhhR
Shewanella oneidensis PhhA/B Hpd HmgA/B/C
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus PhhA/B Hpd HmgA/B/C TyrR
0.0125
0.0250
0.1
Haemophilus ducreyi
0.0250
H. actinomycetemcomitans
Haemophilus somnus
Haemophilus influenzae
Pasteurella multocida TyrR
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
Erwinia herbicola
Citrobacter braakii
Pantoea agglomerans
Salmonella typhimurium
Escherichia coli
Shigella flexneri TyrR

Figure 7



Break in the lines, designated by, ‖, represents a distance of 0.1

Figure 8

PROTEINS: Structure, Function, and Bioinformatics 55:656–677 (2004)

Modeling Structurally Variable Regions in Homologous Proteins With Rosetta

Carol A. Rohl,[1]* Charlie E.M. Strauss,[2] Dylan Chivian,[3] and David Baker[3]
[1]*Department of Biomolecular Engineering, University of California, Santa Cruz, California*
[2]*Biosciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico*
[3]*Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, Washington*

ABSTRACT **A major limitation of current comparative modeling methods is the accuracy with which regions that are structurally divergent from homologues of known structure can be modeled. Because structural differences between homologous proteins are responsible for variations in protein function and specificity, the ability to model these differences has important functional consequences. Although existing methods can provide reasonably accurate models of short loop regions, modeling longer structurally divergent regions is an unsolved problem. Here we describe a method based on the de novo structure prediction algorithm, Rosetta, for predicting conformations of structurally divergent regions in comparative models. Initial conformations for short segments are selected from the protein structure database, whereas longer segments are built up by using three- and nine-residue fragments drawn from the database and combined by using the Rosetta algorithm. A gap closure term in the potential in combination with modified Newton's method for gradient descent minimization is used to ensure continuity of the peptide backbone. Conformations of variable regions are refined in the context of a fixed template structure using Monte Carlo minimization together with rapid repacking of side-chains to iteratively optimize backbone torsion angles and side-chain rotamers. For short loops, mean accuracies of 0.69, 1.45, and 3.62 Å are obtained for 4, 8, and 12 residue loops, respectively. In addition, the method can provide reasonable models of conformations of longer protein segments: predicted conformations of 3Å root-mean-square deviation or better were obtained for 5 of 10 examples of segments ranging from 13 to 34 residues. In combination with a sequence alignment algorithm, this method generates complete, ungapped models of protein structures, including regions both similar to and divergent from a homologous structure. This combined method was used to make predictions for 28 protein domains in the Critical Assessment of Protein Structure 4 (CASP 4) and 59 domains in CASP 5, where the method ranked highly among comparative modeling and fold recognition methods. Model accuracy in these blind predictions is dominated by alignment quality, but in the context of accurate alignments, long protein segments can be accurately modeled. Notably, the method correctly predicted the local structure of a 39-residue insertion into a TIM barrel in CASP 5 target T0186. Proteins 2004;55:656–677.**

Key words: comparative protein structure modeling; homology modeling; fragment assembly; CASP; loop modeling; structurally variable region

INTRODUCTION

Comparative modeling is based on the observation that proteins with similar sequences almost always share similar structures (for review, see Ref. 1). Structure prediction by comparative modeling is initiated by aligning the query sequence to a parent sequence of known structure. For residues that can be aligned, the backbone coordinates of the model are based closely on the coordinates of the parent structure. Residues in the query sequence that cannot be aligned to the parent sequence because of insertions and deletions cannot, by definition, be modeled by using the parent structure as a template. Models for such segments of the protein must be constructed by alternate prediction methods. In addition, regions where sequence similarity is weak and/or alignment uncertain are also candidates for methods targeted at predicting conformations for protein segments corresponding to alignment gaps. Currently, ~30% of known sequences have sufficient sequence similarity to a known structure for current comparative modeling methods. One third of these sequences are similar over <80% of their length; consequently, complete three-dimensional (3D) models cannot be generated by homology-based methods alone.[2] Because sequence and structure divergence between homologous family members is responsible for changes in protein function and specificity, accurately modeling the struc-

Grant sponsor: Howard Hughes Medical Institute; Grant sponsor: Interdisciplinary Training in Genomic Sciences; Grant number: T32 HG00035-06.

*Correspondence to: Carol A. Rohl, Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064. E-mail: rohl@ucsc.edu

Received 8 October 2003; Accepted 14 July 2003

Published online 1 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10629

tural differences between similar structures is an important goal of protein structure prediction.

Traditionally, loop modeling is defined as the problem of constructing 3D atomic models for short protein segments corresponding to loops on the proteins surface that connect regular secondary structure elements. Much attention has been focused on this problem, and several methods have been described that predict loop conformations up to about 8–12 residues with accuracies comparable to the accuracy of models obtained by homology-based methods. Modeling of longer segments of protein structures has received significantly less attention and remains generally an unsolved problem. Here, we use the term structurally variable region (SVR) modeling to refer to prediction of the conformation of any protein segment in the context of a framework or template structure, regardless of the segment length, secondary structure content, or surface exposure. We describe a method based on the successful de novo structure prediction method Rosetta[3,4] for modeling SVRs. In combination with an alignment algorithm to generate template structures, the SVR modeling method allows complete atomic models of proteins to be generated by combining both homology-based and de novo strategies. Complete models for comparative modeling and fold recognition targets were predicted by using this combined strategy and submitted to the Critical Assessment of Structure Prediction (CASP 4 and CASP 5), where the method was ranked highly.[5,6]

A thorough review of the extensive literature on loop modeling methodologies is beyond the scope of this article. Instead, we focus here on general approaches to and distinctions among loop modeling strategies to place the Rosetta-based strategy in context with respect to other methods, as well as to highlight novel contributions. Loop modeling methods primarily differ in the method of conformation generation and in the evaluation or scoring of alternate conformations. Algorithms can be generally grouped into knowledge-based methods, de novo or ab initio strategies, and combined approaches. The knowledge-based approach uses the database of experimental protein structures as a source of loop conformations.[7–14] Generally, such loop conformations are evaluated by using a knowledge-based potential or rule-based filters, evaluating such criteria as geometric fit and sequence similarity to select likely loop conformations. In the de novo approach, loop conformations are generated by a variety of methods including molecular dynamics,[15,16] simulated annealing,[17,18] exhaustive enumeration or heuristic sampling of a discrete set of (ϕ, ψ) angles,[19–23] random tweak,[24,25] or analytical methods.[26,27] Such de novo generated conformations are often evaluated by using components of molecular mechanics force fields, with a variety of treatments of electrostatics and solvation.[18,25,28] Knowledge-based potentials have also been used in combination with conformational sampling methods, as have energy functions that combine molecular mechanics force-field terms with statistical potentials.[17]

Several studies have also combined knowledge-based and de novo methods in a hybrid approach to loop modeling. Mas et al.[29] used a combination of database and conformational search methods to model the hypervariable loops in an antibody; the conformational search method was used both to verify conformations selected from the set of canonical structures and to model de novo the conformation of one loop for which canonical conformations could not be reliably identified. Martin et al.[30] proposed a method that relied on conformational search for short (<5 residues) loops and database search for medium (6–7 residues) loops; long loops were predicted by a hybrid approach in which the central residues in a database-selected conformation were reconstructed by conformational search. In a sequential approach to combining database and conformational searching, VanVlijmen and Karplus[11] demonstrated that performance of a database method could be improved by subsequent optimization and ranking with a molecular mechanics potential. Deane and Blundell[13] described a combined approach that uses the consensus predictions of a knowledge-based and de novo loop modeling method. Sudarsanam et al.[31] used exhaustive sampling of dimers of discrete set of (ϕ, ψ) angles but derived this angle set from angles sampled in known protein structures.

The Rosetta-based method described here is a novel approach to combining database-derived conformations and de novo prediction for loop modeling. In the Rosetta method, originally developed for de novo prediction of entire protein domains, structures sampled by local sequences are approximated by the distribution of structures seen for those short sequences and related sequences in the Protein Data Bank (PDB). These fragments are then assembled in a Monte Carlo search strategy using a scoring function that favors nonlocal properties of native protein structures such as hydrophobic burial, compactness, and pairing of β-strands. Using only primary sequence information, successful de novo Rosetta predictions of entire protein domains yield models on the order of 3–7 Å Cα root-mean-square deviation (RMSD) to native for substantial fragments (>60 residues) of the query sequence.[32,33]

The fragment assembly strategy used by Rosetta is currently perhaps the most successful method for de novo structure prediction, and it may be particularly well suited to modeling SVRs in proteins. By building conformations from smaller fragments, the problem of adequate sampling in the database for longer loops encountered in knowledge-based methods can be potentially overcome, while still restricting the conformational search to a tractable size—a problem encountered by de novo loop modeling methods for longer segments. Furthermore, the fragment buildup strategy allows regular secondary structure to be easily incorporated in predictions for longer SVRs, overcoming a limitation of many de novo loop modeling strategies. Consequently, the method is not limited to protein loops but is applicable to SVRs of any size. A final novel approach used in the current method is the simultaneous modeling of side-chain and backbone conformations using idealized geometry and a rotamer approximation of side-chain conformation. The use of rotamer representations of

the side-chains during optimization of backbone conformations further reduces the complexity of the search space while allowing an atom-based potential function to be used for optimization.

MATERIALS AND METHODS

The SVR modeling method described here uses the Rosetta scoring function and fragment insertion methodologies developed for de novo structure prediction.[3,4] In brief, a customized library of fragments for each three-and nine-residue window in the protein sequence is selected from a database of known protein structures on the basis of local sequence similarity and similarity between the known and predicted secondary structure. These fragments are then assembled by using a Monte Carlo simulated annealing search strategy in which fragments are randomly inserted into the protein chain by replacing the backbone torsion angles in the protein chain with those in the fragment. The resulting protein conformation is then evaluated according to a protein database-derived scoring function that rewards native-like protein properties (see below). In the standard Rosetta protocol for de novo structure prediction, a reduced representation of the protein is used: backbone heavy atoms and Cβ atoms are explicitly included, whereas side-chains are represented by a single centroid. As described below, structure prediction simulations used a combination of this reduced protein representation and an all heavy atom representation with explicit side-chain rotamers.

Database Search

Like the de novo Rosetta protocol, the modeling strategy used here also uses a combination of database-derived fragments that approximate local conformational preferences and a Monte Carlo simulated annealing minimization of a target energy function. Given a sequence alignment between the query and a parent homologue of known structure, the protein structure is divided into template regions and SVRs, which are defined as sections of the chain whose torsion angles cannot be approximated by using those of the parent structure and may include loops, larger insertions, regions of uncertain alignment, and aligned regions where significant structural perturbations are expected. Template regions include all residues whose backbone torsion angles and Cartesian coordinates are taken directly from the parent structure and held fixed throughout the simulations. Cofactors and ligands present in the homologue structure are included in the fixed template coordinates. As in the standard Rosetta protocol, a customized library of three- and nine-residue fragments is selected for the protein sequence and used as described below.

For each SVR of 15 residues or less, an additional customized library of 200–300 possible conformations for the SVR is extracted from the protein structure database. The scoring function used to evaluate these initial loop conformations is a modified form of the scoring function used to generate fragment libraries in the de novo Rosetta protocol and ranks protein segments according to four criteria: 1) sequence profile–profile similarity over the SVR, 2) similarity of the predicted and known secondary structure over the SVR, 3) similarity between secondary structure of template residues adjacent to the SVR in the query and the candidate database conformation, and 4) geometric fit of the database conformation to the template. The process proceeds in two stages. First, a large database representative of the diversity in the nonredundant PDB is coarsely screened for the top 2000 segments that score well by a composite of the four criteria. To select a final set from this pool of 2000, the segments are ranked first by one of the criteria listed above; the top 250 conformations are then reranked by a second criteria, and the top 25 conformations are retained. The culling process is then repeated with use of other criteria. A variety of orders of ranking criteria are used in the culling sequence, and then all the sets are combined into the final library with duplicates removed. The resulting database of initial conformations is comprised of a narrow set of segments when there is a consensus among the methods and a diverse set when there is a lack of consensus, consistent with the philosophy that a diverse set is preferable to a narrower but potentially incorrect set.

Conformational Search

Multiple independent Monte Carlo-simulated annealing optimizations are conducted from different random seeds for each SVR. For each individual simulation, an initial database conformation is selected randomly from the customized library and built onto the fixed template by requiring chain connectivity at either the N- or C-terminal template-SVR junction and allowing discontinuities in the protein backbone at the other junction. The selection of the junction for chain discontinuity is random for each simulation. Initial conformations for SVRs > 15 residues in length are generated by using the standard Rosetta de novo protocol of randomly inserted nine-residue fragments from the customized library into an initially extended protein chain.[3,4] The generation of these initial conformations is conducted in the context of the template but without evaluation of the geometric fit of the variable region to the template.

SVRs greater than seven residues in length are then subjected to Monte Carlo optimization by using a move set of three- and nine- (for SVRs longer than 15 residues) residue fragments. Fragments are either selected randomly from the library or prescreened to bias selection toward fragments that improve the geometric fit of the SVR to the template stems as measured by a gap penalty (see below). Fragment insertions are also combined with a "wobble" operation in which backbone (ϕ, ψ) angles within or adjacent to the fragment insertion site are perturbed to minimize a cost function consisting of the gap penalty and the torsion potential (see below). In addition to fragment insertions, backbone conformations of SVRs are also modified by using random small changes in ρ, ψ angle pairs of individual residues or compensating changes of (ψ_{i-1}, ϕ_i) pairs. These random angle perturbation moves are also combined with the wobble operation. The combination of

TABLE I. Short Loop Reconstruction Results

				Best score[b]					Best RMSD-G		
Protein	Length	Residues	Native score	RMSD-L (Å)	RMSD-G (Å)	Rank[c]	Score	Enrichment	RMSD-L (Å)	RMSD-G (Å)	Score
2act	8	198–205	−66	2.38	3.79	192	−694	2.84	1.42	2.10	−677
2apr	8	76–83	−914	1.06	2.54	226	−930	1.33	0.33	0.53	−912
2fb4	7	H26–H32	−949	1.12	1.79	15	−961	4.80	0.64	0.97	−958
2fbj	7	H100–H106	−1772	0.34	0.98	1	−1744	4.89	0.34	0.98	−1744
3blm	5	131–135	−1191	0.18	0.43	84	−1215	4.89	0.18	0.21	−1200
3dfr	4	20–23	−1215	0.44	0.80	84	−1237	3.20	0.19	0.34	−1215
3dfr	5	89–93	−1215	0.78	0.96	21	−1256	0.71	0.42	0.83	−1234
3dfr	5	120–124	−1215	0.64	0.76	98	−1231	1.69	0.27	0.34	−1181
3grs	7	83–89	−1447	0.61	0.97	23	−1484	6.04	0.29	0.30	−1464
3sgb	9	E199–E211	−1422	0.80	1.10	6	−1393	1.24	0.66	0.90	−1371
5cpa	7	231–237	−824	0.89	1.22	36	−847	3.38	0.54	0.77	−821
8abp	6	203–208	−913	0.44	0.56	35	−949	4.44	0.27	0.31	−933
8tln	7	E32–E38	−1180	2.10	2.62	71	−1220	0.44	0.84	1.24	−1168
8tln	8	E248–E255	−1221	0.75	1.52	11	−1250	5.78	0.42	0.76	−1239

[a]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations in the 15% best scoring population compared with the entire population.
[b]Best-scoring conformation of 500 independent optimizations.
[c]Rank order by RMSD-G of the best-scoring conformation.

TABLE II. Accuracy of 4-, 8- and 12-Residue Segment Predictions[†]

	Best score[a]			Best RMSD-G	
Length	Mean (median) RMSD-L (Å)	Mean (median) RMSD-G (Å)	Mean enrichment[b]	Mean (median) RMSD-L (Å)	Mean (median) RMSD-G (Å)
4	0.42 ± 0.05 (0.31)	0.69 ± 0.06 (0.54)	2.9 ± 0.3	0.21 ± 0.02 (0.18)	0.30 ± 0.03 (0.25)
8	0.97 ± 0.10 (0.79)	1.45 ± 0.14 (1.20)	3.6 ± 0.2	0.50 ± 0.03 (0.47)	0.67 ± 0.05 (0.59)
12	2.23 ± 0.15 (2.29)	3.62 ± 0.31 (3.65)	2.6 ± 0.2	1.28 ± 0.08 (1.30)	1.66 ± 0.10 (1.76)

[†]Reported uncertainties are the standard error of the mean.
[a]Best-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations in the 15% best-scoring population compared with the entire population.

the various types of conformation modification operators is selected so that moves become progressively more local and less globally perturbing during the course of the simulation. The conformational search is conducted by using a Monte Carlo search followed by a two-stage Monte Carlo minimization strategy.[34] In the first stage, a single line minimization along the gradient is conducted for each attempted move, whereas in the second, the variable metric method of Davidon–Fletcher–Powell is used to find the nearest local minimum of the potential energy surface following each initial conformation modification.[35]

Following the optimization using centroid side-chain representations, full-atom coordinates of the side-chains are generated by using a simulated annealing algorithm and a backbone-dependent rotamer library.[36,37] Additional optimization using small backbone torsion angle perturbations and the full-atom potential (see below) is conducted by using the Monte Carlo minimization strategy, iteratively updating the backbone and side-chain conformations. After modification of the backbone torsion angles, side-chain coordinates are updated by adjusting χ angles to their preferred values for the particular rotamer given the new backbone torsion angles. Rotamers at each position in the SVR and spatially adjacent template regions are then updated, in a randomly selected order, by using the rotamer at each position that gives the best energy according to the full-atom potential (see below). At the conclusion of the energy minimization protocol, the side-chains at all positions are completely repacked by using the simulated annealing protocol.

Energy Function

The standard Rosetta potential is derived from a Bayesian treatment of native protein structures and is comprised of two general classes of terms.[3,4] The first class of terms, which describe the probability of a structure independent of sequence, reward native-like arrangements of secondary structure and overall compactness. A second class of terms, describing the probability of a particular sequence given a structure, reward burial of hydrophobic residues and specific pair interactions and penalize van der Waals clashes. For the portions of simulations using reduced side-chain representations, this standard Rosetta potential is modified to include a gap penalty that penalizes chain discontinuities. This gap penalty is calculated as the RMSD between the fixed coordinates of the first

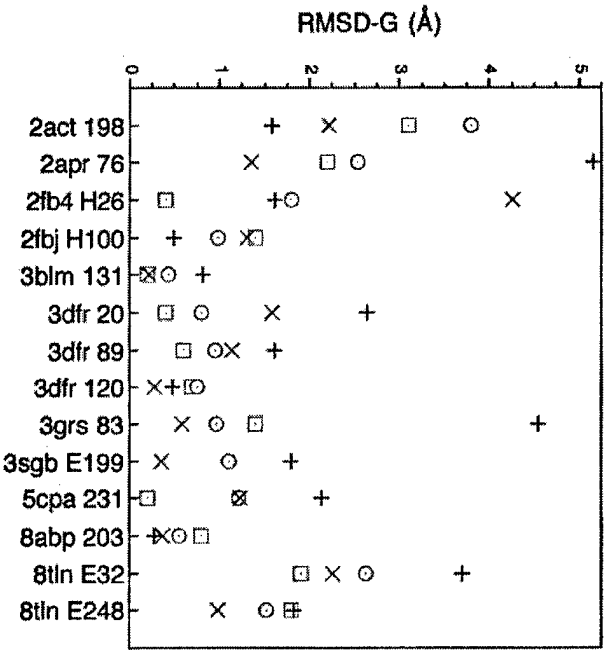


Fig. 1. Comparison of accuracies for loops in Table I predicted by four different methods. The accuracies of previously published predictions by three methods are shown as black plus symbols (Van Vlijmen and Karplus[11]), green x symbols (Fiser et al.[17]), and red squares (Deane and Blundell[13]). Prediction accuracies from the current work are shown as blue circles. RMSD-G is the RMSD of loop residues after superposition of the stem residues (see text). For the Fiser et al. and Rosetta predictions, all backbone heavy atoms (N, CA, C, O) are included in RMSD-G evaluations. For the predictions of Van Vlijmen and Karplus and Deane and Blundell, only N, CA, and C atoms are included in RMSD-G evaluations. Modeled segments are identified by PDB code and first residue.

template residue at each discontinuous template-variable segment junction and the coordinates of this residue determined from the dihedral angles and coordinates of the adjacent variable residue. This same gap penalty score is used in evaluating geometric fit of database conformations to the template.

For all backbone conformation modifications that introduce (ϕ, ψ) angles not taken directly from the fragment library (i.e., random angle perturbation, "wobble" operations, and minimization), torsion angles are evaluated by using a secondary structure-dependent torsion potential.[38] This knowledge-based potential is derived from a nonredundant set of protein X-ray structures of >2.5 Å resolution. For each of the 20 amino acid types in each of three secondary structure types (helix, strand, and other as defined by DSSP[39]), the frequency of (ϕ, ψ) pairs was determined for 10° × 10° bins. Probability distributions were smoothed by using pseudocounts, and the potential was calculated by taking the logarithm of the interpolated probabilities. Randomly selected small angle perturbations, which move backbone conformations away from those represented in the fragment libraries, are discarded according to a Boltzman criterion if they represent an increase in this torsion energy. For moves involving perturbation of backbone angles to minimize a cost function, the torsion potential was included in the target cost function. Backbone ω angles are only modified by fragment insertion and are not evaluated in the torsion potential.

The rotamer packing and rotamer replacement algorithms use the full-atom potential of Kuhlman and Baker[37] with the following modifications: The hydrogen bond potential term used by Kuhlman and Baker is replaced with a hydrogen bond potential derived from PDB statistics. The energies of main-chain–main-chain, side-chain–side-chain, and main-chain–side-chain hydrogen bonds are estimated as a function of the donor and acceptor hybridization and the geometry of the hydrogen bond based on the observed distribution of these parameters in protein crystal structures.[40] The full-atom potential was also supplemented with the torsion potential and gap penalty that was incorporated into the standard Rosetta potential (see above). The complete full-atom potential is thus comprised of 1) the attractive portion of the 12-6 Lennard–Jones potential, 2) a linear repulsive term used in place of the repulsive portion of the 12-6 potential, 3) backbone-dependent internal free energies of the rotamers estimated from PDB statistics, 4) solvation energies calculated by using the model of Lazaridis and Karplus,[41] 5) a knowledge-based pair potential, 6) the hydrogen-bonding potential described above, 7) the knowledge-based backbone torsion potential described above, and 8) the gap penalty described above. This potential is used both for iterative optimization of the SVR backbone and all rotamers and to rank the final population of conformations.

CASP Predictions

For CASP predictions, alignments between the query and parent homologue sequences were generated by using a Smith–Waterman algorithm using PSI-BLAST[42] profile–profile scores, similarity of predicted and known secondary structure, and structural and functional constraints implied by FSSP/DALI topological family sequence profiles.[43] Penalties for insertions and deletions were assigned in a structure-dependent manner using known protein structures to assess the probability of an insertion or deletion of a particular length given the spatial and geometric constraints imposed by flanking residues in the parent structure.[44] Given the alignment between the target sequence and a homologous parent, gaps, insertions, and regions of low-confidence alignment were treated as SVRs.

All SVRs in the target were simultaneously optimized. From the set of resulting models, conformations for each SVR were ranked independently in the context of the fixed template, discarding any conformations that resulted in knots or large-chain discontinuities and retaining the lowest-energy conformations. Combinations of low-energy conformations for each SVR were then evaluated simultaneously to identify low-energy combinations of conformations for all SVRs. The modeling strategy used for CASP 4 targets was an earlier version of the current method and differs from the method described above in several aspects. The primary differences are as follows: 1) the Monte Carlo plus minimization strategy was not used, and all optimization occurred by Monte Carlo search, 2) optimizations

TABLE III. Four-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1aaj	82–85	FTEA	0.14	0.25	4.58	0.13	0.21
1ads	99–102	LKLD	0.22	0.28	2.44	0.16	0.18
1bam	92–95	PIDV	0.61	1.07	5.82	0.62	1.03
1bgc	40–43	HKLC	0.89	1.03	1.47	0.51	0.55
1cbs	21–24	VLGV	0.12	0.29	3.64	0.16	0.18
1fkf	42–45	RNKP	0.16	0.28	3.51	0.11	0.13
1frd	59–62	DQSD	1.07	1.75	0.40	0.24	0.29
1gpr	123–126	NVPS	0.55	0.97	3.20	0.28	0.35
1iab	100–103	FYHE	0.58	0.75	2.89	0.17	0.28
1mba	97–100	GFGV	0.23	1.01	3.56	0.15	0.22
1nfp	37–40	EDTS	1.20	1.49	1.56	0.53	0.58
1pbe	117–120	GATT	0.29	0.57	3.47	0.19	0.25
1pda	139–142	RRPD	0.23	0.32	2.71	0.15	0.17
1pgs	226–229	LGAL	0.83	1.38	0.76	0.17	0.28
1plc	74–77	LSNK	0.37	0.44	3.47	0.26	0.27
1ppn	42–45	TGNL	0.19	0.23	1.29	0.15	0.19
1prn	66–69	GNAA	0.26	0.39	2.98	0.21	0.24
1rcf	111–114	QRGG	0.16	0.25	2.67	0.16	0.25
1tca	287–290	AGPK	0.27	0.43	1.11	0.17	0.22
1thw	194–197	PGSS	1.09	1.28	0.53	0.18	0.28
1tib	46–49	KADA	1.18	1.38	1.24	0.10	0.16
1tml	42–45	FAHH	0.36	0.50	4.22	0.36	0.50
1tys	131–134	SAWN	0.67	1.15	3.07	0.21	0.56
1xif	82–85	TGMK	0.30	0.41	1.87	0.14	0.19
1xnb	30–33	WSNT	0.20	0.51	5.20	0.39	0.44
2cmd	163–166	GKQP	0.28	0.60	2.00	0.19	0.22
2cy3	101–104	KDKK	0.33	0.55	2.53	0.16	0.25
2cyp	127–130	RCGR	0.47	0.81	2.18	0.20	0.33
2cyr[c]	69–71	HAK	0.23	1.12	3.07	0.16	0.45
2exo	161–164	DPTA	0.48	1.03	4.67	0.38	0.40
2sga	44–47	LGFN	0.33	0.43	5.24	0.15	0.25
2sil	220–223	LPSG	0.32	0.66	1.16	0.19	0.26
2tgi	72–75	ASAS	0.34	0.50	5.02	0.15	0.19
3cla	27–30	HRLP	0.13	0.39	0.58	0.11	0.25
4enl	335–338	EKKA	0.25	0.54	6.62	0.17	0.28
4gcr	116–119	FHLT	0.41	0.58	2.49	0.15	0.21
5fd1	81–84	ITEK	0.21	0.53	0.62	0.21	0.31
5p21	75–78	GEGF	0.46	0.87	1.51	0.16	0.28
7rsa	47–50	VHES	0.11	0.18	5.02	0.12	0.17
8abp	55–58	ASGA	0.13	0.20	3.82	0.12	0.16

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.
[c]Three residues only; conformation A of Lys 71 was used as native reference.

generally used only centroid representations of side-chains, although complete heavy atom side-chain coordinates were generated for the final models using the simulated annealing rotamer-packing algorithm, and 3) coding errors present at the time of CASP 4 limited the effectiveness of the optimization. CASP 5 targets used the standard protocol described here, but final loop conformations were selected manually from the top ranked conformations (ranked by energy or cluster size in single-linkage cluster analysis) to eliminate loop combinations resulting in models with steric clashes and/or knots. In addition, although homologous proteins were excluded from the structure database for segment reconstruction tests, homologous proteins were used when available for CASP predictions.

Evaluation of Model Accuracy

To evaluate both the accuracy of the SVR itself, as well as the accuracy of the SVR orientation with respect to the rest of the protein, we report two metrics of model accuracy. RMSD-L is a measure of the model accuracy in a local context and is the RMSD between the model and native over all backbone heavy atoms in the SVR after optimal superposition of the SVR. RMSD-G reports the correctness of both the predicted SVR conformation and its orientation with respect to the template and is the RMSD between the

TABLE IV. Eight-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1351	84–91	LSSDITAS	1.37	1.63	2.44	0.59	0.66
1alc	34–41	SGYDTQAI	0.79	1.09	5.29	0.30	0.46
1art	88–95	FGKGSALI	2.08	3.16	3.82	1.00	1.46
1btl	50–57	DLNSGKIL	0.43	0.63	6.04	0.27	0.41
1cbs	55–62	STTVRTTE	0.52	0.76	4.40	0.35	0.50
1clc	313–320	FRPYDPQY	0.29	1.01	6.27	0.38	0.50
1ddt	127–134	FGDGASRV	2.10	2.94	1.02	0.44	0.69
1fnd	262–269	LKKDNTYV	0.36	0.51	4.22	0.25	0.28
1gky	72–79	QFSGNYYG	0.38	0.72	5.64	0.41	0.48
1gof	606–613	VPSDSGVA	0.76	1.11	3.07	0.54	0.60
1hbq	31–38	DPEGLFLQ	1.17	2.37	1.91	1.21	1.32
1hfc	142–149	SNVTPLTF	0.56	0.68	1.33	0.46	0.49
1iab	48–55	RTTESDYV	2.11	2.92	2.18	0.77	0.83
1ivd	413–420	EGKSCINR	0.97	1.36	4.84	0.51	0.65
1lst	101–108	PIQPTLES	0.47	1.02	2.93	0.35	0.54
1mpp	74–81	TYGTGGAN	1.57	2.55	2.76	0.67	0.73
1nar	192–199	FSNQQKPV	1.04	1.27	4.71	0.50	0.73
1oyc	80–87	GGYDNAPG	0.60	0.68	6.09	0.41	0.51
1phf	85–92	CPFIPREA	0.71	1.12	2.31	0.71	1.12
1poa	71–78	CSQGTLTC	1.15	1.80	4.40	0.50	0.89
1prn	150–157	DPDQTVDS	2.38	2.76	3.20	0.63	0.69
1sbp	107–114	KQIHDWND	0.32	1.07	3.87	0.30	0.45
1thw	18–25	SKGDAALD	0.62	1.01	0.67	0.62	1.01
1tml	187–194	NTSNYRWT	0.75	1.51	3.78	0.37	0.50
1tys	83–90	WADENGDL	0.46	0.86	2.22	0.37	0.47
1xnb	99–106	KSDGGTYD	0.34	0.72	3.87	0.24	0.39
2ayh	123–130	YTNGVGGH	1.32	1.61	3.29	0.31	0.37
2cmd	270–277	LGKNGVEE	1.49	2.55	6.62	0.64	0.95
2ctc	89–96	DYGQDPSF	0.89	1.35	4.53	0.87	1.18
2dri	161–168	PADFDRIK	1.19	1.40	3.60	0.33	0.51
2exo	262–269	MQVTRCQG	0.35	0.41	1.96	0.35	0.41
2ran	26–33	MKGLGTDE	2.33	3.26	2.67	0.90	1.17
2sga	32–43	TTGGSRCS	0.88	1.41	4.93	0.48	0.65
2sns	17–24	AIDGDTVK	0.49	0.59	4.98	0.51	0.57
3cox	109–116	GRGVGGGS	0.78	0.84	3.42	0.38	0.44
3grs	424–431	ANKEEKVV	1.62	3.20	2.84	0.41	0.49
4enl	24–31	TTEKGVFR	0.85	1.43	1.29	0.47	0.69
4fxn	88–95	YGWGDGKW	1.61	1.66	4.18	0.59	1.22
5p21	45–52	VIDGETCL	0.28	0.45	3.47	0.19	0.26
8dfr	65–72	RPLKDRIN	0.41	0.67	3.64	0.51	0.63

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.

model and native of all heavy backbone atoms in the SVR after optimal superposition of three adjacent stem residues on each side of the SVR. For short loops, RMSD-G is the critical measure of accuracy. Most interactions of atoms in short loops are with the template portion of the protein, and correctly predicting the orientation of the loop with respect to the protein core is the primary goal of modeling. For longer SVRs, including insertions comprising intact structural modules, RMSD-L becomes an increasingly relevant metric of model accuracy. Although correct prediction of both the structure of the segment itself and its orientation with respect to the protein core is the end goal of SVR modeling, this goal is generally beyond the capabilities of current methods. Consequently, the accuracy with which SVR structure can be predicted without requiring correct global orientation is a relevant quality indicator. Furthermore, models with correct structure but incorrect orientation likely still include useful structural information.

For purposes of evaluating SVR modeling in CASP targets, a third metric, RMSD-E, is also evaluated to quantify the structural accuracy of the local environment in which the SVR is predicted. RMSD-E is the RMSD between the model and native conformations evaluated over the three stem residues N- and C-terminally adjacent to the SVR after optimal superposition of these residues. For the segment reconstruction tests, the "template" corresponds exactly to the native protein backbone structure, and all RMSD-E values are 0 Å. For CASP targets and, in fact any realistic comparative modeling problem, both

TABLE V. Twelve-Residue Segment Reconstruction Predictions

Protein	Residues	Sequence	Best score[a]		Enrichment[b]	Best RMSG	
			RMSD-L	RMSD-G		RMSD-L	RMSD-G
1541	153–164	NVRSYARMDIGT	0.99	1.51	2.89	0.67	0.97
1arp	201–212	LDSTPQVFDTQF	0.49	0.77	0.93	0.60	0.76
1ctm	9–12	YENPREATGRIV	3.81	5.64	2.00	1.16	2.31
1dts	41–52	SGSEKTPEGLRN	1.58	4.97	1.02	1.33	2.52
1eco	35–46	MAKFTQFAGKDL	3.13	4.15	2.53	0.64	0.94
1ede	150–161	CLMTDPVTQPAF	0.86	0.89	4.84	0.86	0.89
1ezm	122–133	FGDGATMFYPLV	2.06	4.46	3.11	2.06	2.18
1hfc	165–176	RGDHRDNSPFDG	2.33	3.80	1.20	1.79	2.46
1ivd	365–376	TISKDLRSGYET	2.72	4.23	1.38	1.04	1.26
1msc	9–20	LVDNGGTGDVTV	2.75	9.18	2.84	1.85	2.43
1onc	23–34	MSTNLFHCKDKN	2.82	4.03	1.11	1.61	1.72
1pbe	129–140	LHDLQGERPYVT	1.83	2.90	3.38	0.76	0.92
1pmy	77–88	KCAPHYMMGMVA	3.03	4.08	4.00	1.28	1.58
1prn	15–26	VEDRGVGLEDTI	3.16	6.44	3.91	1.98	2.31
1rcf	88–99	TGDQIGYADNFQ	2.15	3.60	1.87	1.83	2.04
1rro	17–28	ECQDPDTFEPQK	2.05	2.66	2.00	0.77	1.02
1scs	199–210	IKSPDSHPADGI	1.85	3.17	2.40	1.06	1.18
1srp	311–322	SDVGGLKGNVSI	1.12	1.16	4.00	0.97	1.10
1tca	305–316	AVGKRTCSGIVT	2.42	3.75	3.91	1.65	1.84
1thg	127–138	WIYGGAFVYGSS	2.89	4.17	2.04	1.89	2.29
1thw	178–189	PDAFSYVLDKPT	2.28	2.83	0.58	1.52	2.09
1tib	99–110	EINDICSGCRGH	2.69	3.12	1.73	0.78	0.94
1tml	243–254	STTNTGDPMIDA	2.97	5.80	3.64	1.75	2.19
1xif	203–214	IERLERPELYGV	1.34	1.64	3.78	0.64	1.08
2cpl	145–156	FGSRNGKTSKKI	3.64	7.45	1.07	1.79	2.07
2cyp	191–202	WGAANNVFTNEF	2.18	2.84	2.13	1.61	2.29
2ebn	136–147	YQTPPPSGFVTP	2.56	3.28	2.09	0.64	0.94
2exo	293–304	LVWDASYAKKPA	1.00	1.51	0.84	0.66	0.96
2pgd	361–372	WRGGCIIRSVFL	2.61	4.32	2.44	1.44	2.04
2rn2	90–101	WKTADKKPVKNV	4.23	7.09	5.47	1.55	2.26
2sil	255–266	ETKDFGKTWTEF	0.53	0.68	5.64	0.53	0.68
2sns	111–122	VAYVYKPNNTHE	1.89	3.14	4.40	2.37	3.02
2tgi	48–59	CPYLWSSDTQHS	2.19	2.86	4.13	1.72	1.96
3b5c	12–23	IQKHNNSKSTWL	3.05	5.22	2.27	0.87	1.04
3cla	176–187	AKYQQEGDRLLL	1.20	1.49	4.49	1.20	1.49
3cox	478–489	VPGNVGVNPFVT	1.65	1.97	1.38	1.26	1.60
3hsc	72–93	RLIGRRFDDAVV	0.55	0.70	2.53	0.51	0.64
451c	16–27	HAIDTKMVGPAY	3.47	5.59	1.51	1.75	2.53
4enl	372–383	SHRSGETEDTFI	2.30	3.69	1.96	1.36	1.79
4ilb	46–57	FVQGEESNDKIP	2.92	4.02	2.04	1.48	2.20

[a]Top-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-G conformations found in the 15% best-scoring population compared with the entire population.

alignment errors and template perturbations contribute to the accuracy of the template from which SVRs are modeled, and these modeling errors result in non-zero RMSD-E values. RMSD-E measures the accuracy only of the stem residues sequentially adjacent to the SVR and does not reflect the structural accuracy of other spatially adjacent residues. Consequently, small RMSD-E values for regions modeled as SVRs in homology models indicate only that the local geometry constraining the ends of the SVR is approximately correct.

RESULTS

The SVR modeling method described here is intended to comprise part of a complete modeling strategy for structure prediction by comparative modeling and fold recognition and was, in fact, applied in combination with an alignment algorithm to generate complete models for all targets in CASP 4 and CASP 5 for which a homologous protein of known structure could be identified. The double-blind CASP experiment offers a realistic test of comparative modeling methods because both alignment errors and structural deviations between a query sequence and the parent structure degrade the accuracy of the local environment in which SVRs must be modeled. However, the blind evaluation of CASP targets is conducted without knowledge of which portions of the model were generated by alignment and which were modeled as structurally divergent. To supplement the analysis of model quality pro-

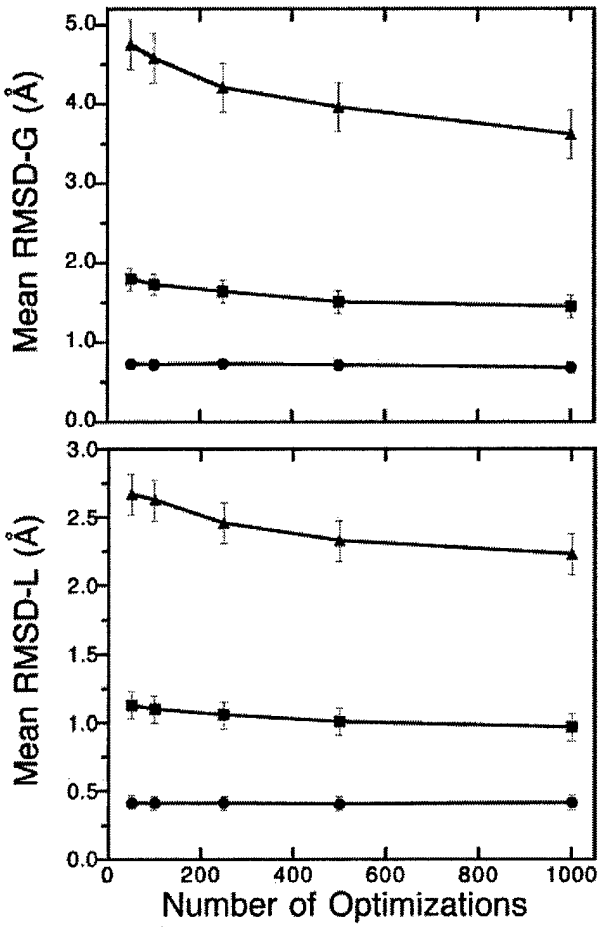


Fig. 2. Mean prediction accuracy as a function of number of independent optimizations.

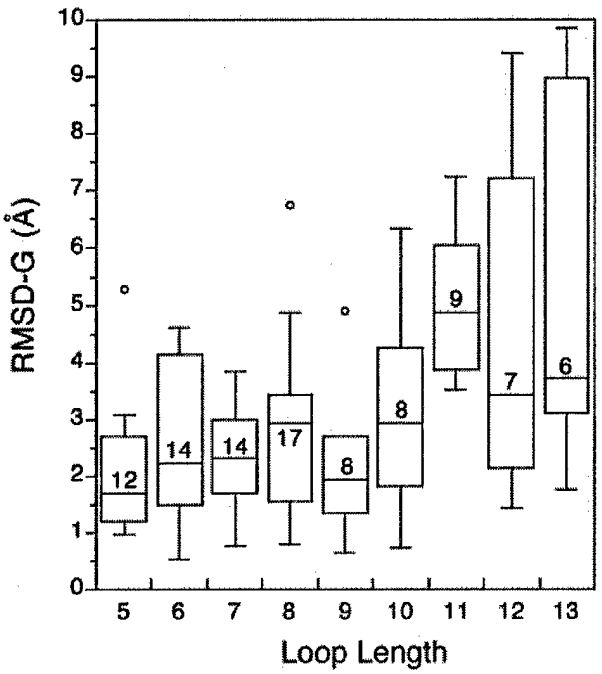


Fig. 3. Box plots of distributions of RMSD-G values for SVRs of lengths 5–13 residues in CASP 5 targets. Only SVRs modeled in the context of reasonably accurate environments (RMSD-E < 1.5Å; see Materials and Methods) are included in the figure. The number in each box indicates the number of modeled SVRs contributing to each distribution.

vided by the CASP assessors and to assess the performance of the Rosetta-based method in the context of realistic modeling errors, we report here the accuracy of the CASP 4 and CASP 5 predictions specifically for segments modeled as SVRs. Complete lists of all regions modeled as SVRs in both CASP 4 and CASP 5 targets for which structures have been released, along with the template and prediction accuracies, are reported here.

In addition to blind CASP predictions of SVRs made in the context of realistic modeling errors, we also present results of predictions in which segments of proteins of known structure are reconstructed in the context of exact templates. Segment reconstruction, although artificial in the sense that it does not represent a realistic structure prediction problem, does allow the SVR method to be assessed in the absence of propagated errors resulting from incorrect alignment and template perturbation. In addition, segment reconstruction has been used as a standard method for assessment of loop modeling methods and allows direct comparison of different modeling strategies. Notably, in the segment reconstruction predictions here, none of the native side-chain conformations are retained; instead, all side-chains are replaced by using the simulated anneal rotamer-packing algorithm. Consequently, although the template backbone is exact, the template side-chain conformations are not, making the segment reconstruction test somewhat more realistic.

Prediction of Short Protein Loops

Results of segment reconstruction predictions made for sets of surface-exposed protein loops, selected and previously predicted by other authors, are given in Tables I and II. The fourteen loops in Table I, varying in length from four to nine residues, are provided as representative examples of predictions for short to medium loops. Several other groups have made predictions for these same segments, allowing direct comparison of several methods on identical examples (Fig. 1). Table II summarizes results obtained for 40 loops each of 4, 8, and 12 residues. Results for all individual predictions in these sets are given in Tables III–V.

For short loops, the Rosetta method effectively samples low RMSD-G conformations. For 38 of 44 loops in the 4- to 5-residue range, conformations <0.5 Å RMSD-G are sampled; in the 7- to 9-residue range, conformations <1 Å are sampled in 40 of 49 cases; and for 30 of 40 12-residue loops, conformations < 2.2 Å are sampled. In most cases, conformations that have energies equal or better than the native loop conformation are sampled (Table I). The effectiveness of the sampling

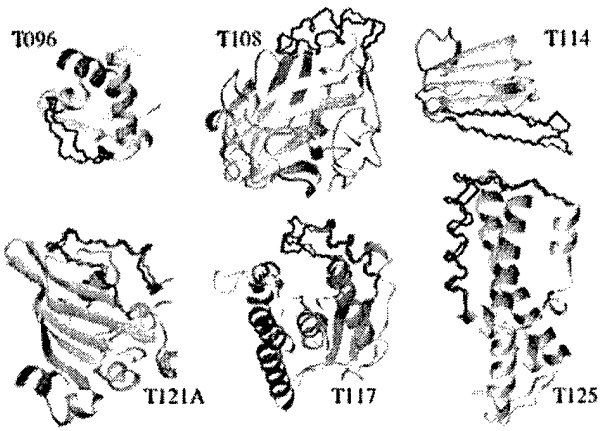


Fig. 4. Top scoring conformations for representative long segment reconstructions. The backbone of the modeled region is shown in blue (native conformation) and red (predicted conformation). The remainder of the backbone structure is shown in gray as a ribbon diagram. Protein structure diagrams were made by using MolMol.[49]

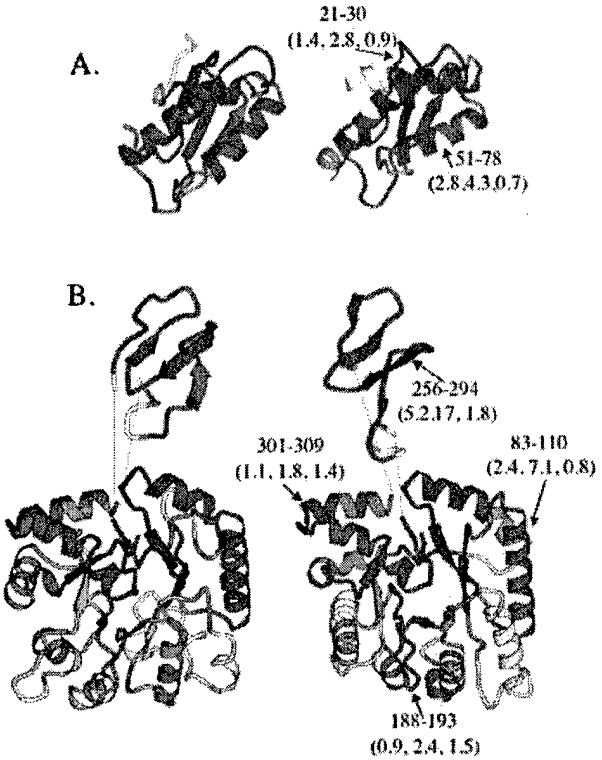


Fig. 6. Selected CASP 5 comparative modeling predictions. Structure diagrams of CASP 5 targets T0130 and T186, residues 44–332, are shown in panels A and B, respectively. The experimental structure is shown on the left, and the first-ranked model is on the right. Regions colored in shades of blue were modeled by using coordinates of a homologue of known structure, whereas regions in shades of orange were modeled as SVRs. For each target, an optimal subset of superimposable residues was found by using the LGA algorithm.[46] Given this structural superposition, the CA deviation between the model and native structure at each position is indicated by color intensity. Regions in dark orange/dark blue have CA deviations of <2 Å after superposition; regions in medium orange/medium blue have CA deviations between 2 and 4 Å, and regions in pale orange/pale blue have CA deviations > 4 Å. Residues are colored identically in the predicted model and experimental structure diagrams. For T0186 (**B**), residues 256–294 have been independently superimposed by using the LGA algorithm. The dotted lines indicate the stem regions to which the SVR termini are connected. Selected SVRs, indicated by arrows, are identified by residue number. Prediction accuracies for these SVRs are given in parenthesis (RMSD-L, RMSD-G, RMSD-E). See text for details. Protein structure diagrams were generated by using Molscript.[50]

method is further illustrated by examining the mean prediction accuracy as a function of the number of independent optimizations conducted (Fig. 2). An increase in mean prediction accuracy on doubling the number of optimizations from 500 to 1000 is seen only for the 12-residue loops. For short loops, the accuracy of prediction is generally limited by discrimination, although ranking of conformations by the potential function does result in significant enrichment (Tables I and II).

Although accurate predictions are made in the context of the native protein, significantly poorer performance is seen for short loop modeling in CASP targets where local template geometries are less than perfect. In CASP 5, 59 domains were modeled by using homology to a protein of known structure. In the targets for which structures are available, 215 regions of ≤13 residues were modeled as SVRs. Of these, 177 are nonterminal segments, with template-imposed geometric constraints similar to those of the segments reconstructed in native proteins. Ninety-seven of these SVRs were modeled in the context of reasonably accurate local templates (RMSD-E < 1.5 Å). The distribution of prediction accuracies for loops meeting these criteria are shown in Figure 3. The mean accuracies of loop predictions are significantly worse than those seen in the segment reconstruction tests, indicating, as noted by many previous authors, that the accuracy of loop modeling in real comparative modeling applications is determined almost entirely by alignment accuracy and template distortions. In addition, loop modeling in real homology models is complicated by the fact that multiple, potentially interacting, loops must be modeled within the same structure.

Prediction of Long SVRs

A motivating goal in developing Rosetta for SVR modeling is to provide a modeling method that is not limited only to short loops but is also applicable to predicting longer insertions and structural differences between homologous proteins. To examine the accuracy of the Rosetta method in predicting conformations of longer SVRs, 10 segments ranging from 13 to 34 residues were selected from CASP 4 comparative modeling targets to be reconstructed in the context of the native protein. For each of the proteins, the region of greatest structural divergence with respect to the closest structural match in the PDB, as determined by the CASP 4 assessors,[45] was selected as the segment to be reconstructed. Unlike the shorter protein loops discussed above, these segments do not necessarily correspond to surface-exposed protein loops. Results for these 10 predictions are given in Table VI, and structures of low-energy conformations for some successful predictions are given in Figure 4. For these longer protein segments, the accuracy

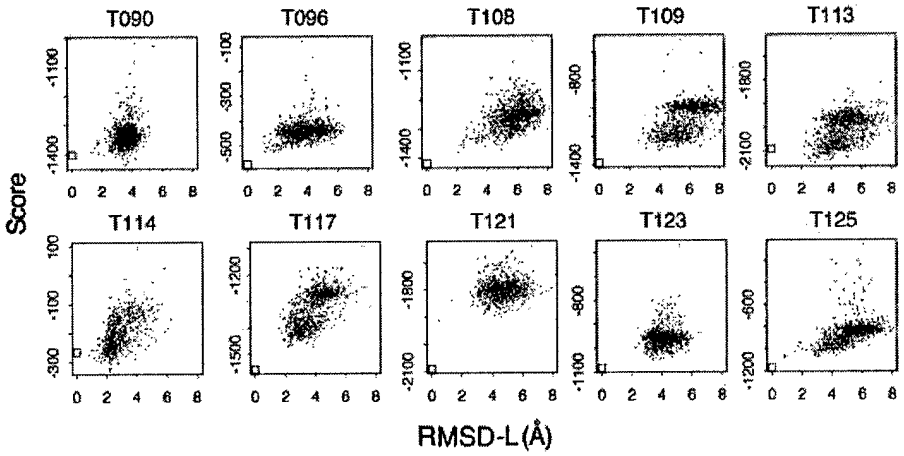


Fig. 5. Conformation discrimination for long SVR reconstructions. The correlation between the final score and RMSD-L is shown for independent optimizations of each predicted segment in Table III. The score of the native segment conformation in each case is indicated by the open square.

TABLE VI. Long Segment Reconstruction Results

				Best score[a]					Best RMSD-L		
Protein	Residues	Length	Native score	RMSD-L (Å)	RMSD-G (Å)	Rank[c]	Score	Enrichment[b]	RMSD-L (Å)	RMSD-G (Å)	Score
T090	77–91	15	−1402	3.54	6.11	474	−1434	1.6	1.33	3.45	−1394
T096	19–31	13	−572	1.21	2.42	3	−521	1.8	1.12	2.82	−460
T108	139–155	17	−1452	2.05	2.84	1	−1415	2.8	2.05	2.84	−1415
T109	48–81	34	−1393	4.00	20.4	49	−1302	2.7	2.62	9.72	−1242
T113	203–223	21	−2092	2.91	3.91	19	−2145	3.0	2.21	2.62	−2080
T114	51–65	15	−264	2.22	3.08	163	−325	3.1	0.84	1.28	−275
T117	138–159	22	−1557	2.19	2.39	24	−1471	2.4	1.60	3.93	−1401
T121	65–82	18	−2089	0.51	0.88	1	−1913	1.5	0.51	0.88	−1913
T123	28–41	14	−1082	3.88	6.95	360	−1047	1.2	2.29	3.78	−996
T125	94–118	25	−1165	0.84	2.48	2	−1076	4.3	0.82	2.23	−1058

[a]Best-scoring conformation of 1000 independent optimizations.
[b]Ratio of the relative occurrence of the 15% lowest RMSD-L conformations in the 15% best-scoring population compared with the entire population.
[c]Rank order by RMSD-L of the best-scoring conformation.

of the predictions is limited both by the conformational search and by discrimination. Native structures show significantly better scores than all sampled conformations in 7 of the 10 examples. In most cases, some correlation between the accuracy of the predicted segment conformations and the evaluated scores is observed (Fig. 5), with an average enrichment of 2.5 ± 0.9 (Table VI), suggesting that additional sampling might result in improved prediction accuracies.

The Rosetta method was also used to make predictions for long SVRs in CASP 4 and CASP 5 targets. Fifty SVRs ranging in length from 14 to 78 residues were predicted in CASP 4 targets for which structures are available, and 74 SVRs ranging in length from 14 to 123 residues were predicted in CASP 5 targets for which structures are available. Table VII gives results of the long SVR predictions in CASP 5 targets that were modeled in the most accurate local template environments (RMSD-E < 2.5Å) and the identity and prediction accuracies for all SVRs in all CASP 4 and CASP 5 targets are given in Tables VIII and IX. As with short loops, performance on long SVRs degrades significantly in the context of realistic modeling errors. In segment reconstruction, 7 of 10 examples have RMSD_L < 3 Å and 5 have RMSD-G < 3 Å. Of 32 long SVRs in CASP 5 targets (Table VII), 12 have RMSD-L < 3 Å, and only 2 have RMSD-G < 3 Å. As noted above (see Materials and Methods), RMSD-E only measures the correctness of stem geometry, not the overall accuracy of the environment. Because longer segments generally have more nonlocal contacts than short, surface-exposed loops, RMSD-E significantly underestimates the true environment error for long SVR predictions. Consequently, examining predictions that have correct local structures, even in the absence of correct orientation is warranted. However, it is important to note that many of the predictions with best local accuracy correspond to single regular secondary structure elements (e.g., a single helix in a TIM barrel that was modeled as an SVR because of alignment uncertain-

TABLE VII. Long SVR Predictions in CASP 5 Targets†

Target	Region	Length	RMSD-L (Å)	RMSD-G (Å)	RMSD-E (Å)	End-to-end distance (Å)
T0147	7–20	14	3.85	5.42	1.90	16.3
T0168	298–311	14	3.51	7.29	0.44	8.5
T0149	19–33	15	3.18	6.33	1.35	15.3
T0168	249–263	15	3.54	6.47	1.06	8.0
T0168	279–293	15	4.23	12.15	1.78	4.5
T0169	124–138	15	5.53	12.46	0.85	19.9
T0184	108–122	15	4.25	7.77	0.39	12.6
T0185	176–190	15	4.61	7.68	0.98	17.3
T0186	197–211	15	4.42	7.75	1.37	10.6
T0134	161–176	16	3.18	6.30	2.11	22.1
T0151	84–99	16	0.73	1.46	0.56	5.8
T0154	15–30	16	0.59	1.89	0.36	20.8
T0185	248–263	16	2.07	4.13	1.57	13.4
T0195	58–73	16	3.07	8.02	2.49	21.0
T0165	224–240	17	3.90	8.42	2.17	13.5
T0168	222–238	17	2.14	4.72	2.49	9.3
T0183	96–112	17	0.68	2.39	0.99	21.7
T0184	35–51	17	4.40	8.16	0.92	14.6
T0186	116–132	17	2.63	11.08	2.28	12.4
T0189	16–33	18	4.67	12.38	1.21	4.6
T0193	149–166	18	1.04	3.40	0.56	15.7
T0160	94–112	19	2.30	6.52	0.92	6.2
T0172	56–75	20	1.97	3.08	0.42	11.6
T0133	228–251	24	0.87	1.18	0.41	12.7
T0141	86–111	26	6.42	18.99	2.01	15.2
T0149	98–124	27	3.04	5.39	0.48	12.3
T0130	51–78	28	2.81	4.34	0.65	8.6
T0142	45–72	28	3.41	4.59	0.47	19.7
T0186	83–110	28	2.38	7.12	0.78	11.0
T0165	120–150	31	7.45	12.82	1.23	13.0
T0195	91–124	34	6.58	20.45	1.84	11.4
T0186	256–294	39	5.20	17.27	1.71	9.3

†Predictions for SVRs of length 14 and greater submitted as part of first-ranked models in CASP 5. Only predictions made in the context of the most accurate local environments (RMSD-E < 2.5Å) are included in the table.

ties). The end-to-end distance of each SVR in the native protein is reported to help identify those SVRs whose conformations are highly constrained by stem locations.

Despite the difficulty in drawing general conclusions from SVRs in CASP targets, these predictions illustrate the promise of the method for long SVR modeling. Examples from two CASP 5 targets are shown in Figure 6. The template portion of T0130 was generated by alignment to 1fbaA [blue region in Fig. 6(A)]. The two proteins are 23% identical over the structurally superimposable portions, permitting a reasonably accurate alignment to be obtained. Relative to the optimal structural superposition of 1fbaA and experimental T0130 structure, the alignment in the CASP 5 model is 76% accurate and 29% complete. By intent, our alignment algorithm was biased for high specificity at the expense of sensitivity, and we relied on SVR modeling with Rosetta to complete the models. Two internal segments of T0130 were modeled as SVRs: residues 21–30 comprise the C-terminus of the first helix, the N-terminus of the first strand, and the intervening loop; residues 51–78 comprise the second helix and the two long loops connecting this helix to the sheet. Both of these loops are among the best predictions made for loops of their size in CASP 5 targets [Fig. 6(A)].

The template portion of T0186, residues 44–332, was generated by alignment to 1gkpA. The two proteins are 15% identical over structurally superimposable regions, and the alignment used to generate the template is only 46% accurate and 50% complete with respect to the structural superposition. Despite significant alignment errors, four SVRs were modeled in the context of reasonably accurate stem geometries (RMSD-E $\leq$ 1.8 Å). Residues 83–100 comprise one helix on the surface of the TIM barrel along with the connecting loops; residues 188–193 are a loop connecting a helix-strand pair, and residues 301–309 are a loop connecting a helix–helix pair on one end of the barrel. As in T0130, these three SVR predictions are among the most accurate predictions for SVRs of their size in the CASP 5 targets. In addition, when the entire protein model is compared with the experimental structure without concern for the modeling method used, these three SVRs, as well as the two internal SVRs in T0130 discussed above, are of approximately the same accuracy as regions of the model generated by accurate alignment (see Fig. 6).

TABLE VIII. All SVR Predictions In First-Ranked Models of CASP 4 Targets

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0089	1–10	10	1.41	4.07	2.98
T0089	26–31	6	2.55	5.26	3.12
T0089	46–54	9	3.94	10.39	0.25
T0089	64–93	30	10.35	19.83	0.10
T0089	119–159	41	11.16	20.95	0.28
T0089	166–170	5	2.01	4.64	1.73
T0089	198–206	9	2.61	4.47	1.19
T0089	223–230	8	1.94	3.49	1.16
T0089	249–253	5	0.56	1.36	7.25
T0089	263–290	28	4.68	7.69	0.26
T0089	312–331	20	3.70	6.78	0.43
T0089	359–419	61	6.88	16.68	5.97
T0090	1–57	57	13.94	26.44	0.86
T0090	66–70	5	0.58	1.61	1.26
T0090	78–91	14	4.35	14.36	3.99
T0090	149–158	10	2.05	10.17	1.24
T0090	177–209	33	4.55	21.44	2.01
T0092	1–38	38	4.82	20.58	0.45
T0092	51–57	7	1.47	2.97	2.03
T0092	74–82	9	2.56	4.35	0.14
T0092	98–111	14	2.43	5.83	2.31
T0092	116–127	12	1.97	4.14	0.09
T0092	132–144	13	2.37	2.67	0.63
T0092	162–210	49	5.51	26.03	3.20
T0092	218–222	5	0.66	1.47	0.89
T0092	229–234	6	2.66	7.43	2.01
T0096	1–9	9	1.67	4.71	1.09
T0096	21–34	14	4.02	8.14	0.67
T0096	41–46	6	0.45	0.62	0.58
T0096	64–70	7	2.74	4.97	7.82
T0100	25–44	20	8.85	16.36	0.09
T0100	54–58	5	1.48	5.57	1.99
T0100	69–73	5	0.88	1.76	2.18
T0100	75–78	4	1.87	2.48	0.78
T0100	92–114	23	5.94	8.05	1.99
T0100	118–121	4	0.91	1.92	0.54
T0100	131–154	24	3.94	6.69	3.94
T0100	158–166	9	2.50	3.61	3.26
T0100	176–179	4	2.02	3.46	2.02
T0100	182–186	5	1.56	2.14	0.89
T0100	196–199	4	1.68	4.36	2.86
T0100	202–206	5	1.58	2.79	2.16
T0100	216–232	17	3.73	7.69	4.76
T0100	253–262	10	2.83	4.42	3.55
T0100	266–287	22	3.82	13.42	2.78
T0100	299–315	17	5.85	12.50	4.00
T0100	319–323	5	1.91	5.53	2.14
T0100	332–352	21	4.98	24.82	2.63
T0100	361–366	6	2.80	8.65	1.30
T0101	26–47	22	5.69	34.03	1.75
T0101	57–64	8	2.45	6.16	2.70
T0101	70–75	6	1.81	4.34	2.77
T0101	84–94	11	1.93	6.32	0.67
T0101	96–105	10	4.81	8.87	2.31
T0101	117–134	18	3.67	4.35	2.01
T0101	150–178	29	6.52	23.77	1.03
T0101	181–190	10	1.43	2.00	0.76
T0101	196–232	37	8.47	25.72	0.90
T0101	239–244	6	1.61	2.02	1.98
T0101	260–283	24	9.71	16.07	7.95
T0101	300–307	8	2.12	3.08	1.94
T0101	317–328	12	3.77	9.20	3.97

TABLE VIII. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0101	342–362	21	5.26	20.54	4.65
T0101	419–425	7	2.71	9.82	1.60
T0103	1–23	23	3.73	17.70	0.50
T0103	34–36	3	1.23	1.52	2.24
T0103	54–58	5	3.09	5.20	1.04
T0103	66–71	6	1.75	8.55	2.01
T0103	124–128	5	1.63	1.94	1.06
T0103	137–147	11	3.23	3.98	3.26
T0103	172–187	16	5.36	13.04	0.69
T0103	204–213	10	4.22	8.17	0.25
T0103	229–285	57	15.77	17.49	0.37
T0103	318–372	55	10.18	21.59	0.14
T0108	1–39	39	3.58	18.76	1.35
T0108	44–48	5	1.44	3.37	3.55
T0108	53–59	7	1.84	6.27	1.19
T0108	72–86	15	5.79	9.22	1.48
T0108	94–103	10	2.66	2.95	2.73
T0108	147–158	12	2.41	13.81	1.45
T0108	164–174	11	3.31	5.84	0.46
T0108	190–196	7	2.04	5.34	1.12
T0109	1–8	8	1.88	9.59	2.60
T0109	32–44	13	4.48	11.02	0.15
T0109	50–85	36	4.75	13.61	5.45
T0109	118–129	12	1.92	6.40	2.70
T0109	134–158	25	6.29	13.43	3.25
T0109	177–182	6	3.47	9.16	0.31
T0111	1–1	1	0.63	2.04	0.72
T0111	30–33	4	2.09	2.88	0.50
T0111	79–85	7	1.25	1.94	0.64
T0111	139–142	4	1.81	3.78	2.63
T0111	199–203	5	0.46	1.04	3.51
T0111	234–239	6	0.61	1.23	10.19
T0111	261–267	7	1.64	4.20	4.68
T0111	306–310	5	0.58	0.85	0.40
T0112	11–15	5	2.00	3.62	2.86
T0112	48–53	6	2.03	5.39	1.39
T0112	113–122	10	3.79	7.61	0.29
T0112	151–154	4	0.17	1.28	2.61
T0112	160–165	6	0.68	1.92	1.94
T0112	190–194	5	0.34	0.84	0.95
T0112	212–216	5	2.02	2.87	2.16
T0112	220–228	9	3.23	6.16	0.76
T0112	261–264	4	0.98	1.10	1.43
T0112	273–283	11	2.68	4.44	0.77
T0112	336–342	7	2.31	4.00	2.18
T0112	349–352	4	0.61	3.92	1.67
T0113	1–12	12	1.84	5.84	4.65
T0113	96–110	15	4.79	6.14	3.97
T0113	137–146	10	3.37	8.72	2.63
T0113	202–227	26	2.83	4.94	2.14
T0113	241–247	7	0.99	1.25	0.81
T0113	256–261	6	2.29	15.73	0.21
T0114	1–15	15	5.70	24.34	1.25
T0114	59–62	4	2.44	5.77	0.46
T0114	70–72	3	1.58	3.72	0.22
T0115	1–4	4	0.97	3.92	2.81
T0115	9–13	5	1.28	5.27	0.73
T0115	29–95	67	10.47	20.40	1.30
T0115	136–168	33	8.69	13.72	2.60
T0115	181–186	6	1.98	10.69	4.02
T0115	194–222	29	9.44	26.52	1.70
T0116	1–18	18	6.31	45.22	0.13
T0116	43–59	17	7.46	12.99	2.54

TABLE VIII. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0116	72–82	11	3.54	5.48	0.78
T0116	104–117	14	4.86	9.14	4.70
T0116	124–130	7	3.29	4.33	2.77
T0116	136–151	16	3.73	10.57	1.14
T0116	158–164	7	1.63	2.37	0.64
T0116	168–174	7	2.61	8.28	0.84
T0116	180–252	73	2.49	21.74	0.46
T0117	1–23	23	1.56	4.82	1.99
T0117	36–46	11	1.97	3.44	1.51
T0117	71–78	8	2.04	2.88	1.54
T0117	89–101	13	3.77	5.90	1.94
T0117	135–146	12	1.30	3.17	0.90
T0117	173–176	4	1.88	3.91	1.64
T0117	191–200	10	3.11	7.43	1.60
T121	1–3	3	N/A	N/A	N/A
T0121	67–76	10	3.89	5.94	1.56
T0121	102–112	11	0.61	1.07	2.78
T0121	132–136	5	2.16	4.69	0.62
T0121	188–191	4	1.94	2.64	3.94
T0122	1–2	2	0.76	6.87	4.41
T0122	26–33	8	0.77	1.05	14.68
T0122	77–81	5	1.06	1.66	3.00
T0122	173–180	8	1.42	5.10	4.00
T0122	241–248	8	2.63	4.48	1.03
T0125	1–10	10	1.85	3.04	4.76
T0125	32–43	12	1.12	1.67	0.15
T0125	67–83	17	3.08	5.09	6.51
T0125	99–117	19	3.75	5.89	7.95
T0125	135–141	7	1.31	8.13	1.00
T0127	1–23	23	2.56	7.25	0.23
T0127	41–47	7	1.60	2.71	1.85
T0127	68–145	78	13.14	14.26	7.23
T0127	153–161	9	0.59	1.39	0.62
T0127	170–185	16	2.84	9.34	4.74
T0128	1–12	12	0.62	2.39	1.95
T0128	66–72	7	1.73	4.39	2.02
T0128	147–151	5	0.52	1.89	0.87
T0128	212–222	11	4.01	11.02	0.86

[a]Not adjusted for missing density in experimental PDB files. Superposition and RMSD calculations use only atoms for which density is reported in the experimental PDB file.
[b]Root-mean-square deviation of residues in the SVR following optimal superposition of the SVR residues.
[c]Root-mean-square deviation of residues in the SVR following optimal superposition of the three stem residues N- and C-terminally adjacent to the SVR.
[d]Root-mean-square deviation of the three stem residues N- and C-terminally adjacent to the SVR following optimal superposition of these stem residues.

The fourth SVR in T0186, residues 256–294, is a small subdomain inserted into the TIM barrel. Although the relative orientation of this SVR was not predicted correctly (RMSD-G = 17Å), the local structure of four-stranded meander is correctly predicted with an RMSD-L of 5.2Å (Fig. 6). If the distortions at the SVR termini are disregarded, the local RMSD significantly improves: a sequence-dependent iterative superposition with a 4 Å cutoff using the LGA algorithm[46] yields an optimal fragment match of 30 residues with an RMSD-L of 2.4 Å. Notably, the prediction of this SVR by the Rosetta-based method was significantly better than any other submitted prediction.

DISCUSSION

The Rosetta-based method for SVR modeling represents a new approach to combining database and de novo strategies for modeling protein segments, both short loops and longer SVRs. The assembly of conformations from smaller fragments allows the benefits of database methods and de novo loop modeling methods to be combined. Iterative optimization of the backbone and side-chain conformations, using a rotamer approximation for side-chains, which to our knowledge has not been previously applied to loop modeling, allows detailed atomic interactions to be evaluated, while significantly restricting the complexity of the conformational search. Allowable confor-

TABLE IX. All SVR Predictions in First-Ranked Models of CASP 5 Targets

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0130	1–13	13	2.98	6.55	0.22
T0130	21–30	10	1.44	2.84	0.90
T0130	51–78	28	2.81	4.34	0.65
T0130	81–114	34	5.02	9.34	0.37
T0132	1–18	18	7.33	15.51	0.56
T0132	52–57	6	3.45	5.11	2.58
T0132	101–112	12	2.76	7.21	0.51
T0132	122–130	9	2.34	2.77	1.59
T0132	133–154	22	1.50	3.63	0.38
T0133	1–29	29	2.04	20.19	0.13
T0133	59–66	8	2.43	3.45	1.05
T0133	98–109	12	1.71	3.42	1.42
T0133	117–126	10	3.03	6.33	1.22
T0133	145–179	35	6.18	16.17	3.02
T0133	197–215	19	4.82	8.89	3.49
T0133	228–251	24	0.87	1.18	0.41
T0133	270–279	10	1.15	1.83	1.25
T0133	287–312	26	9.20	18.34	1.57
T0134	878–882	5	0.87	11.92	0.56
T0134	899–905	7	2.00	2.86	1.10
T0134	928–943	11	3.44	5.17	0.98
T0134	966–976	12	1.15	3.40	0.54
T0134	982–993	3	0.41	1.29	1.74
T0134	1003–1005	13	5.01	7.97	4.66
T0134	1020–1032	16	3.18	6.30	2.11
T0134	1038–1053	5	2.46	6.78	4.29
T0134	1060–1064	7	4.08	8.52	5.39
T0134	1070–1076	6	2.00	3.15	2.76
T0134	1082–1087	7	0.46	0.90	0.23
T0137	41–49	9	0.80	1.36	0.44
T0137	97–102	6	2.21	4.20	0.61
T0137	108–112	5	0.26	2.03	0.41
T0137	119–123	5	0.30	1.06	0.25
T0138	1–4	4	1.94	5.28	0.52
T0138	46–53	8	2.25	2.94	1.11
T0138	58–63	6	1.69	4.13	1.37
T0138	84–89	6	2.55	9.55	3.86
T0138	96–103	8	3.26	7.12	4.09
T0138	106–116	11	2.20	3.21	1.65
T0138	132–135	4	1.24	7.76	1.42
T0141	1–30	30	8.06	22.10	0.31
T0141	55–75	21	4.25	9.05	3.37
T0141	86–111	26	6.42	18.99	2.01
T0141	118–128	11	2.86	3.87	0.69
T0141	144–150	7	2.75	4.07	2.12
T0141	154–171	18	5.27	10.23	2.62
T0141	175–187	13	4.00	13.20	0.14
T0142	1–8	8	1.60	2.35	0.16
T0142	45–72	28	3.41	4.59	0.47
T0142	91–103	13	3.10	3.87	2.05
T0142	106–114	9	2.73	3.19	1.70
T0142	136–144	9	2.66	6.58	2.85
T0142	155–164	10	3.13	5.64	2.10
T0142	200–208	9	1.31	1.73	0.49
T0142	234–239	6	0.86	1.46	0.86
T0142	248–257	10	2.76	3.48	0.92
T0142	262–269	8	2.26	2.73	0.97
T0142	279–282	4	0.35	4.27	0.18
T0147	1–3	3	0.80	3.95	0.69
T0147	7–20	14	3.85	5.42	1.90

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0147	38–52	15	2.66	5.79	2.65
T0147	62–68	7	2.93	6.16	2.56
T0147	72–82	11	2.93	9.76	2.53
T0147	90–94	5	1.12	2.70	2.21
T0147	98–116	19	4.47	10.77	2.71
T0147	121–127	7	1.90	3.37	3.31
T0147	131–141	11	3.67	5.41	2.63
T0147	149–154	6	1.57	3.29	1.63
T0147	158–175	18	2.54	5.29	2.08
T0147	182–186	5	0.64	2.52	1.61
T0147	190–202	13	4.40	8.07	2.12
T0147	210–216	7	2.01	4.39	2.66
T0147	219–245	27	5.01	14.47	1.75
T0149	1–5	5	0.81	7.72	0.21
T0149	19–33	15	3.18	6.33	1.35
T0149	37–43	7	2.95	5.19	3.53
T0149	58–77	20	6.36	20.28	4.95
T0149	84–94	11	3.59	5.16	2.74
T0149	98–124	27	3.04	5.39	0.48
T0149	148–154	7	2.65	4.66	2.21
T0149	174–184	11	3.01	4.56	1.96
T0149	186–193	8	4.11	7.12	4.35
T0149	195–318	124	15.98	33.44	1.76
T0150	–2–6	8	3.03	4.88	0.07
T0150	94–100	7	0.37	2.25	0.38
T0151	1–6	6	2.10	3.04	0.34
T0151	21–28	8	2.02	3.08	0.76
T0151	36–52	17	1.84	2.64	0.75
T0151	84–99	16	0.73	1.46	0.56
T0151	103–164	62	6.35	9.32	0.19
T0153	30–35	6	1.11	1.28	0.85
T0153	52–58	7	0.42	0.77	0.56
T0153	95–103	9	1.10	2.14	0.49
T0153	119–154	36	6.60	24.15	0.26
T0154	1–11	11	3.67	18.61	0.12
T0154	15–30	16	0.59	1.89	0.36
T0154	54–62	9	1.83	2.36	0.32
T0154	110–117	8	2.57	3.43	0.77
T0154	241–248	8	2.26	3.52	1.82
T0154	254–266	13	3.93	9.87	0.44
T0154	286–309	24	2.13	9.09	0.37
T0155	84–91	8	0.43	1.13	0.37
T0155	119–133	15	0.69	2.82	0.67
T0157	1–2	2	0.47	4.60	0.51
T0157	21–26	6	0.46	2.37	1.73
T0157	35–42	8	1.71	6.74	1.21
T0157	59–71	13	2.95	4.97	1.94
T0157	95–121	27	4.46	5.61	1.70
T0157	133–138	6	0.45	1.06	0.21
T0159	1–8	8	2.30	11.74	0.31
T0159	12–18	7	2.01	2.54	2.55
T0159	31–40	10	3.37	4.58	2.96
T0159	54–59	6	1.43	3.31	1.72
T0159	63–73	11	3.43	4.02	1.40
T0159	77–88	12	3.15	5.46	2.11
T0159	103–111	9	3.53	6.84	1.58
T0159	114–124	11	1.55	7.51	2.50
T0159	146–153	8	2.74	5.61	3.26
T0159	186–193	8	2.37	6.04	1.86
T0159	211–229	19	5.79	15.98	2.55
T0159	265–282	18	3.95	14.68	4.28
T0159	291–296	6	2.11	3.97	1.73
T0159	298–309	12	1.39	11.66	1.46

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0160	−3–7	10	3.84	8.91	0.22
T0160	76–84	9	3.27	4.91	0.65
T0160	94–112	19	2.30	6.52	0.92
T0165	1–54	54	18.78	41.91	0.29
T0165	64–68	5	1.02	1.35	0.80
T0165	75–82	8	2.17	4.88	0.80
T0165	87–91	5	1.05	1.27	1.29
T0165	95–101	7	2.74	3.53	3.20
T0165	105–111	7	0.40	1.24	1.98
T0165	120–150	31	7.45	12.82	1.23
T0165	167–171	5	1.36	3.08	1.27
T0165	189–199	11	1.79	3.63	0.37
T0165	206–212	7	1.84	2.98	0.79
T0165	224–240	17	3.90	8.42	2.17
T0165	252–260	9	1.01	2.70	0.76
T0165	284–289	6	1.94	2.55	0.43
T0165	298–304	7	0.78	1.15	0.46
T0167	1–4	4	2.05	3.38	0.12
T0167	111–123	13	3.01	3.11	0.28
T0167	127–146	20	5.66	6.82	0.58
T0167	183–185	3	1.21	4.77	0.61
T0168	56–59	4	0.50	1.40	0.51
T0168	63–69	7	2.83	3.84	1.08
T0168	91–129	39	11.76	22.63	5.98
T0168	150–164	15	5.36	10.73	6.24
T0168	196–209	14	5.40	10.41	9.40
T0168	222–238	17	2.14	4.72	2.49
T0168	249–263	15	3.54	6.47	1.06
T0168	271–275	5	1.94	3.46	2.20
T0168	279–293	15	4.23	12.15	1.78
T0168	298–311	14	3.51	7.29	0.44
T0168	323–327	5	1.65	11.86	0.28
T0169	5–10	6	0.43	1.49	0.92
T0169	23–28	6	1.41	7.53	2.90
T0169	36–42	7	1.17	1.74	1.36
T0169	62–67	6	2.65	3.71	0.56
T0169	110–115	6	2.23	4.60	0.75
T0169	124–138	15	5.53	12.46	0.85
T0172	1–7	7	2.43	13.43	1.71
T0172	19–27	9	2.85	4.62	3.06
T0172	45–49	5	0.49	0.99	0.50
T0172	56–75	20	1.97	3.08	0.42
T0172	80–85	6	0.85	1.43	2.00
T0172	107–218	112	13.48	17.59	3.33
T0172	245–249	5	2.06	3.55	2.00
T0172	264–282	19	6.25	9.20	4.53
T0172	293–299	7	0.53	4.96	1.17
T0182	1–5	5	0.77	2.16	0.12
T0182	47–52	6	0.31	0.52	0.51
T0182	249–250	2	0.55	1.86	0.73
T0183	1–26	26	4.79	43.44	0.60
T0183	40–47	8	0.49	0.78	0.31
T0183	56–63	8	0.45	1.15	0.36
T0183	78–84	7	1.36	1.70	0.57
T0183	96–112	17	0.68	2.39	0.99
T0183	142–148	7	0.73	1.18	0.25
T0183	155–166	12	0.47	1.44	0.43
T0183	183–189	7	1.98	2.54	0.38
T0183	197–206	10	0.53	0.74	0.59
T0183	221–229	9	0.51	0.64	0.40
T0183	235–248	14	1.21	15.27	0.34
T0184	1–9	9	0.48	1.74	0.20
T0184	35–51	17	4.40	8.16	0.92
T0184	70–79	10	1.11	1.87	0.74

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0184	108–122	15	4.25	7.77	0.39
T0184	139–146	8	1.13	1.57	0.81
T0184	162–172	11	4.21	6.92	4.26
T0184	178–185	8	1.66	2.23	0.75
T0184	193–199	7	0.95	2.40	0.71
T0184	207–212	6	1.55	4.07	0.37
T0184	235–240	6	0.83	1.87	0.14
T0185	1–3	3	0.33	1.83	0.22
T0185	15–27	13	2.65	3.75	2.46
T0185	51–60	10	1.73	3.01	0.39
T0185	88–103	16	2.04	4.93	2.67
T0185	127–133	7	2.39	3.09	2.00
T0185	160–172	13	2.91	3.65	0.56
T0185	176–190	15	4.61	7.68	0.98
T0185	217–221	5	1.78	2.48	0.60
T0185	236–243	8	3.08	8.28	2.77
T0185	248–263	16	2.07	4.13	1.57
T0185	306–312	7	0.62	2.23	0.32
T0185	317–322	6	0.82	1.94	0.38
T0185	329–341	13	0.83	1.77	0.93
T0185	346–369	24	2.39	3.47	2.15
T0185	374–407	34	4.92	11.48	1.49
T0185	416–421	6	2.16	4.15	0.78
T0185	426–433	8	2.75	3.40	1.44
T0185	443–457	15	0.26	0.99	0.14
T0186	10–14	5	1.29	2.61	0.50
T0186	27–34	8	2.86	6.19	4.00
T0186	52–63	12	4.39	9.40	0.79
T0186	83–110	28	2.38	7.12	0.78
T0186	116–132	17	2.63	11.08	2.28
T0186	150–159	10	2.63	5.49	2.64
T0186	177–181	5	2.52	3.93	1.97
T0186	188–193	6	0.85	2.36	1.51
T0186	197–211	15	4.42	7.75	1.37
T0186	230–238	9	3.46	8.29	3.26
T0186	244–250	7	2.12	3.33	1.35
T0186	256–294	39	5.20	17.27	1.71
T0186	301–309	9	1.12	1.73	1.44
T0186	327–331	5	0.94	1.35	0.35
T0186	344–351	8	1.02	5.11	2.06
T0186	354–359	6	1.59	2.46	1.96
T0188	6–16	11	2.87	6.06	1.34
T0188	46–56	11	2.88	4.87	1.11
T0189	1–3	3	0.30	0.65	0.25
T0189	16–33	18	4.67	12.38	1.21
T0189	63–68	6	1.00	1.63	0.56
T0189	74–81	8	1.36	2.94	1.13
T0189	88–95	8	2.85	3.80	3.26
T0189	101–107	7	2.63	5.57	4.70
T0189	112–124	13	0.79	1.78	1.90
T0189	141–151	11	1.35	3.52	0.75
T0189	174–185	12	3.15	3.97	2.00
T0189	193–202	10	3.72	6.98	1.65
T0189	219–223	5	0.59	1.19	0.72
T0189	229–235	7	1.73	3.00	2.08
T0189	241–250	10	2.33	4.83	2.06
T0189	273–279	7	1.57	2.07	0.54
T0189	299–307	9	2.03	5.71	1.92
T0189	317–319	3	1.33	11.06	1.42
T0190	1–5	5	0.66	1.34	0.21
T0190	29–34	6	0.95	1.83	0.39

TABLE IX. (Continued)

Target	Region[a]	Length	RMSL[b] (Å)	RMSG[c] (Å)	RMSE[d] (Å)
T0190	51–62	12	2.90	3.59	1.30
T0190	90–96	7	2.68	3.55	0.89
T0191	1–105	105	6.64	10.39	0.21
T0191	143–147	5	1.79	5.28	1.06
T0191	164–175	12	2.70	8.64	3.29
T0191	180–190	11	3.27	6.89	1.32
T0191	196–208	13	4.62	8.97	1.18
T0191	215–219	5	1.99	2.69	0.87
T0191	224–234	11	3.52	6.95	1.94
T0191	254–268	15	3.31	14.61	5.43
T0192	1–3	3	0.76	2.52	0.32
T0192	27–36	10	2.45	4.25	1.42
T0192	41–45	5	0.69	1.21	2.04
T0192	47–51	5	2.04	4.07	1.68
T0192	58–70	13	4.14	13.77	2.04
T0192	78–89	12	0.70	2.14	0.26
T0192	143–153	11	4.04	7.25	1.49
T0192	159–171	13	1.45	21.24	0.18
T0193	1–13	13	3.63	5.54	0.24
T0193	22–28	7	1.20	2.14	2.46
T0193	54–60	7	1.06	6.91	2.85
T0193	64–81	18	4.63	7.00	4.35
T0193	98–105	8	2.73	4.40	0.74
T0193	114–125	12	4.49	6.73	1.83
T0193	132–141	10	2.28	4.26	1.61
T0193	149–166	18	1.04	3.40	0.56
T0193	170–178	9	2.14	6.91	3.80
T0193	189–195	7	3.33	8.54	5.64
T0193	199–211	13	3.09	13.25	1.52
T0195	1–12	12	4.92	11.85	0.24
T0195	35–47	13	3.61	3.82	0.71
T0195	58–73	16	3.07	8.02	2.49
T0195	77–79	3	0.85	4.54	2.97
T0195	91–124	34	6.58	20.45	1.84
T0195	142–154	13	4.60	6.77	1.25
T0195	173–180	8	0.96	1.43	0.59
T0195	188–215	28	5.90	11.07	4.59
T0195	217–232	16	4.00	6.43	3.43
T0195	242–253	12	2.23	3.10	1.50
T0195	259–266	8	1.87	2.68	1.39
T0195	291–299	9	3.10	5.62	0.14

[a]Not adjusted for missing density in experimental PDB files. Superposition and RMSD calculations use only atoms for which density is reported in the experimental PDB file.
[b]Root-mean-square deviation of residues in the SVR following optimal superposition of the SVR residues.
[c]Root-mean-square deviation of residues in the SVR following optimal superposition of the three stem residues N- and C-terminally adjacent to the SVR.
[d]Root-mean-square deviation of the three stem residues N- and C-terminally adjacent to the SVR following optimal superposition of these stem residues.

mations for protein segments up to about five or six residues are adequately sampled in known protein structures,[47] and fragment assembly is unlikely to significantly improve the accuracy of predictions for segments below this size. Because accurate backbone conformations can be selected from known structures, however, the benefits of the rotamer approximation for optimizing atomic interactions likely do contribute to the accuracy of the method for such short segments. Conversely, for long SVRs, sampled conformations may not be sufficiently accurate that optimization of detailed atomic interactions can improve the predictions, but fragment assembly is likely to be critical for effective sampling of backbone conformations.

For short loops, the mean prediction accuracies obtained by the Rosetta method are comparable with those obtained by other loop modeling approaches. Among the best results reported are those of Fiser et al.[17] who obtain RMSD-G values of 0.79, 1.89, and 4.24 Å for 4, 8, and 12 residue loops, respectively. Other recent successful methods have reported mean RMSD-G values of 0.85 and 1.45 for five and eight residue loops[24] and 1.00 and 3.09 Å for four- and eight-residue loops.[13] The mean prediction accuracies

obtained here, 0.59, 1.45, and 3.62 Å for 4, 8, and 12 residue loops, are at least comparable with these methods. Given that real loop modeling does not happen in environments of perfect accuracy, it is unclear what significance, if any, the differences in performance of various methods in the segment reconstruction test have for actual loop modeling. Although the mean prediction accuracies of the best methods are reasonably comparable, the most accurate method for any particular loop region varies, as illustrated in Figure 1. In this small sample set, the de novo prediction method of Fiser et al.[17] and the consensus hybrid approach of Deane and Blundell[13] are the most likely to yield the best prediction, whereas the database method of Van Vlijmen and Karplus[11] yields the best prediction in two cases. The Rosetta method gives good predictions on average but does not result in the top ranked prediction in any of these examples. The fact that the Rosetta-based method does not use native side-chain conformation information in segment reconstructions may contribute in part to this ranking.

Although a variety of methods can predict short loop conformations with reasonable accuracy, reliable prediction of the conformation of long SVRs is an unsolved problem. Because the conformational space accessible to a polypeptide chain increases exponentially with increasing chain length, the difficulty of the structure prediction problem increases dramatically as chain length increases and, consequently, the accuracy with which protein segments are predicted decreases. A hypothesis guiding this work is that the fragment buildup strategy used in the Rosetta method could combine the strengths of database methods with conformational search methods. By assembling shorter fragments to generate conformations for longer regions, the conformational database can be extrapolated, allowing longer protein segments to be modeled with greater accuracy. The predictions obtained for 13- to 35-residue segments, although insufficient to give statistically significant estimates of mean accuracies, illustrate that the method is indeed extendable to long SVRs. In 5 of the 10 cases examined, predictions >2.5 Å RMSD-G were obtained for segments ranging from 13 to 34 residues. In addition, examples from CASP 5 comparative modeling targets, although anecdotal, are quite promising. In several cases where long SVRs were modeled in the context of reasonably accurate alignments, regions modeled as SVRs have accuracies comparable with regions modeled by alignment to a homologue of known structure (Fig. 6).

Given these promising results, how can additional improvements in the method be obtained? For longer segments, conformational sampling becomes a limiting factor in the accuracy of predictions. The native conformation is frequently significantly lower in energy than the lowest-energy conformation sampled (Fig. 5), indicating that significant improvement in the accuracy of long segment predictions could be obtained by additional sampling. For short segments, the potential is not sufficiently accurate to identify the native conformation in general (Table I). Although improvements in the potential clearly would be required to improve the accuracy of the short segment predictions, a bigger practical limitation on the accuracy of short segments is the alignment and environment accuracy. Perhaps the most fruitful target for improvements to the method is in the selection of optimal predictions from the population of sampled conformations. The current discrimination scheme relies solely on ranking conformations according to the potential used for optimization. Clustering has been previously shown to improve discrimination in both de novo structure prediction[48] and loop modeling[13,24] by identifying conformations corresponding to wide energy basins. Addition of clustering to the discrimination scheme is likely to yield an improvement in the current method as well.

CONCLUSION

Comparative modeling provides 3D models for proteins based on sequence similarity to a protein of known structure, and improving the accuracy and completeness of such models requires methods capable of modeling structural divergences between homologous proteins. Because the differences between related structures are responsible for differences in functional specificity, the ability to accurately model SVRs in homologous sequences is required to fully exploit comparative models for functional insight. Although both optimization and database search methods are able to provide accurate models for short loop regions in proteins, accurate structural modeling of longer SVRs in proteins is an unsolved problem. Providing accurate models of longer insertions and template perturbations, however, is perhaps the most biologically relevant application of comparative modeling because such structural changes add novel functions and specificities to protein scaffolds. Here we use the fragment buildup strategy of the de novo prediction algorithm Rosetta in an attempt to overcome some of the sampling limitations that restrict the accuracy of modeling methods by extrapolating the structure database to cover longer protein segments. The resulting method performs as well as existing loop modeling methods on short loops, and initial results for longer segments illustrate the promise of the method for predicting structures of long SVRs as well.

ACKNOWLEDGMENTS

CAR was supported by the Interdisciplinary Training in Genomic Sciences program. DC is a fellow of the Program in Mathematics and Molecular Biology at the Florida State University, with funding from the Burroughs Wellcome Fund Interfaces Program.

NOTE ADDED IN PROOF

The Rosetta potential and methods for local sampling and rapid fragment screening used in this study are described in detail in a forthcoming volume of Methods in Enzymology (Rohl CA, Strauss CEM, Misura KMS, Baker D. Meth Enzym 2004;383:66–93).

REFERENCES

1. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.

2. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
3. Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;35:82–95.
4. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
5. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4 Proteins 2001;Suppl 5:22–38.
6. Sippl MJ, Lackner P, Dominguez FS, Prlíc A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl 5:55–67.
7. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
8. Martin ACR, Thornton JM. Structural families in loops of homologous proteins: automatic classification, modeling and application to antibodies. J Mol Biol 1996;263:800–815.
9. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830.
10. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformation class of short and medium size loops connecting regular secondary structures: application to comparative modeling. J Mol Biol 1997;267:352–367.
11. Van Vlijmen WWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;257:975–1001.
12. Wojcik J, Mornon J-P, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289:1469–1490.
13. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 2001;10:599–612.
14. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng 2001;7:473–478.
15. Bruccoleri RE, Karplus M. Conformational sampling using high-temperature molecular dynamics. Biopolymers1990;29:1847–1862.
16. Hornak V, Simmerling C. Generation of accurate protein loop conformations through low-barrier molecular dynamics. Proteins 2003;51:577–590.
17. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2001;9:1753–1773.
18. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized Born model of solvation effects. Proteins 1999;35:173–183.
19. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1986;1:146–163.
20. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments in proteins by systematic search. Biopolymers 1987;26:137–168.
21. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
22. Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small medium and large loops in proteins. Biopolymers 2001;60:153–168.
23. DePristo MA, de Bakker, PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
24. Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 1987;26:2053–2085.
25. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
26. Go N, Scheraga HA. Ring closure and local conformation deformations of chain molecules. Macromolecules 1970;3:178–187.
27. Wedemeyer W, Scheraga HA. Exact analytical loop closure in proteins using polynomial equations. J Comp Chem 1999;20:819–844.
28. de Bakker PIW, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential ant the AMBER force field with the Generalized Born solvation model. Proteins 2003;51:21–40.
29. Mas MT, Smith KC, Yarmush DL, Aisaka K, Fine RM. Modeling the anti-cea antibody combining site by homology and conformational search. Proteins 1992;14:483–498.
30. Martin AC, Cheetham JC, Rees AR. Modeling antibody hypervariable loops: combined algorithm. Proc Natl Acad Sci USA. 1989;86:9268–9272.
31. Sudarsanam S, DuBose RF, March CJ, Srinivasan S. Modeling protein loops using a Φ I+1, Ψ i dimer database. J Mol Biol1995;206:759–777.
32. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in CASP4: Progress in ab initio protein structure prediction. Proteins 2001;Suppl 5:119–126.
33. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: successes, failures and prospects for complete automation. Proteins 2003. Forthcoming.
34. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA 1987;84:6611–6615.
35. Press WH, Teukolski SA, Vetterling WT, Flannery BP. Numerical recipes in Fortran 77: the art of scientific computing, 2nd ed. Cambridge: Cambridge University Press; 2001.
36. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
37. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
38. Bowers PM, Strauss CEM, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR 2000;18:311–318.
39. Kabsch W, Sander C. Dictionary of protein secondary: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
40. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.
41. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
42. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
43. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
44. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CEM, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the ROBETTA server. Proteins 2003. Forthcoming.
45. http://PredictionCenter.llnl.gov/CASP4.
46. Zemla A. LGA program: a method for finding 3-D similarities in protein structures. 2000; accessed at http://PredictionCenterllnl-gov/local/lga
47. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng 1994;7:953–960.
48. Shortle D, Simons KT, Baker D. Clustering of low energy conformations near the native structures of small proteins. Proc Natl Acad Sci 1998;95:11158–11162.
49. Koradi R, Billeter M, Wüthrich K. MOMOL: a program for display and analysis of macromolecular structures. J Mol Graphics 1996;14:51–55.
50. Kraulis P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24:946–950.

Efficient Nucleic Acid Signature Development for Broad Spectrum Pathogen Detection

Jason D. Gans

Los Alamos National Laboratory

Bioscience Division, B5

Los Alamos NATIONAL LABORATORY EST 1943

The World's Greatest Science Protecting America

NNSA

Acknowledgements

Los Alamos National Laboratory
Bioscience Division

Informatics Team
Cathy Cleland
Norman Dogget
Rob Leach
Jian Song
Charlie Strauss
Chris Stubben
Murray Wolinsky*
Yan Xu
Wendy Zheng

Experimental Validation
John Dunbar*
Lance Green
Scott White

Funding
Department of Homeland Security

Los Alamos
NATIONAL LABORATORY
EST. 1943
The World's Greatest Science Protecting America

NNSA

The need for pathogen detection

Los Alamos
NATIONAL LABORATORY
EST 1943
The World's Greatest Science Protecting America

NNSA

Why DNA?

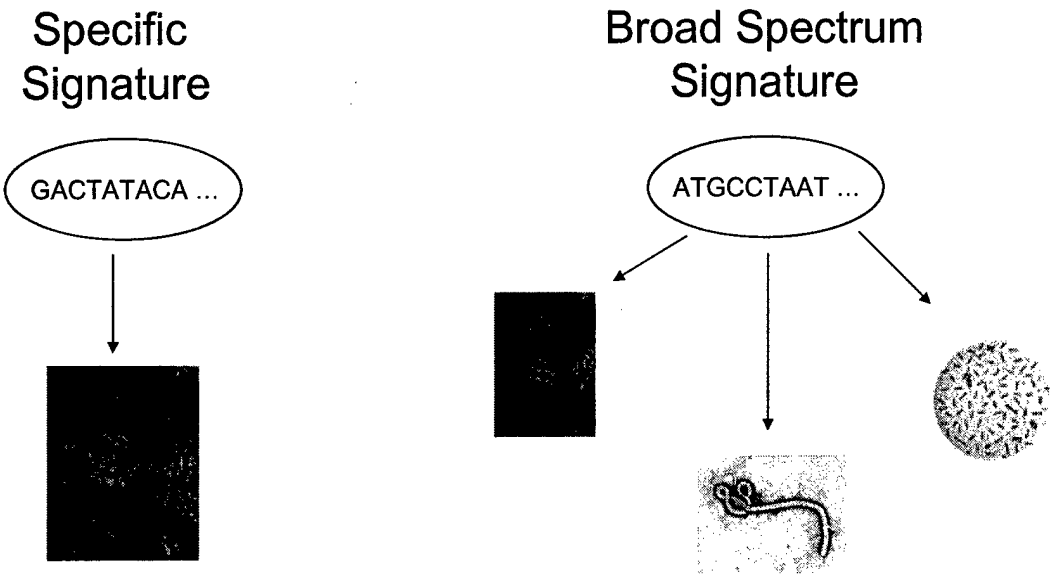
DNA signature development can exploit the growing number of sequenced genomes

- ~ 200 complete bacterial genomes available
- ~ 1700 complete viral genomes available

Mature technologies for high throughput detection

- Gene chip, PCR

Specific vs broad spectrum signatures



Detecting multiple targets with a single experiment reduces cost

Los Alamos NATIONAL LABORATORY EST 1943

The World's Greatest Science Protecting America

Images courtesy of http://www.microbelibrary.org

NNSA

How to define sequence similarity?

The answer depends on how we ask the question

The assay format, e.g.
- PCR amplification
- DNA chip probe hybridization
- Single base extension

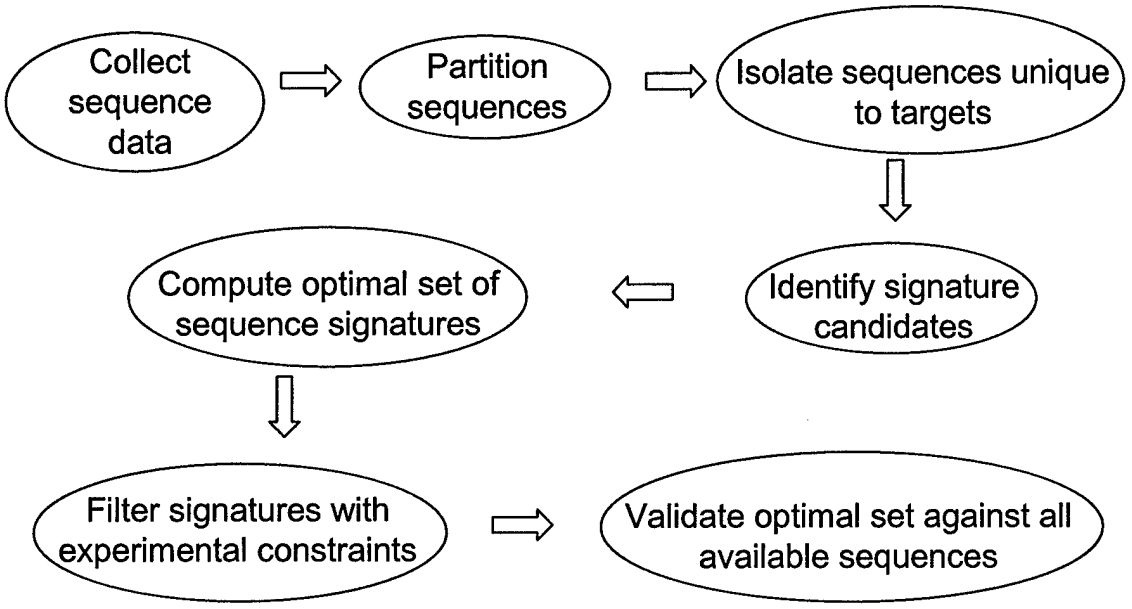
defines similarity.

A sequence similarity metric that supports common assay formats is DNA melting temperature, T_m

$T_m^{A,B} > T_m^0 \Rightarrow$ Sequences A and B are equivalent

$T_m^{A,B} < T_m^0 \Rightarrow$ Sequences A and B are dissimilar

Signature Detection Pipeline



Collect Sequence Data

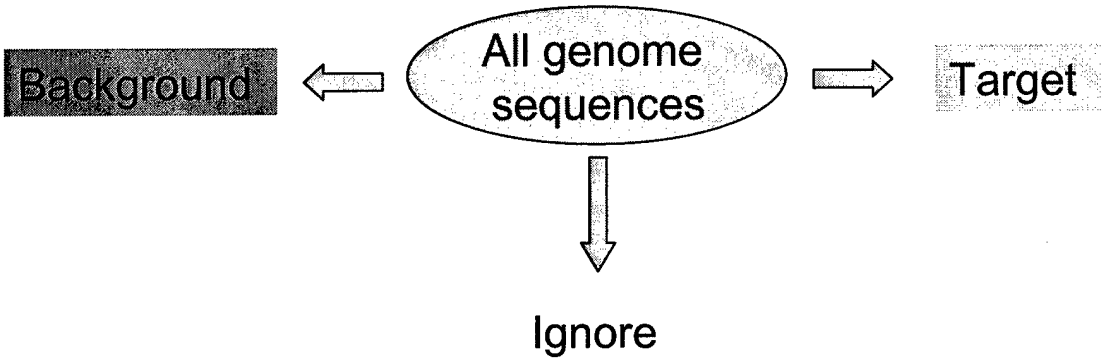
What taxa need to be detected ?

- Food borne
- Air/soil/water borne
- Environmental
- Animal/Plant vector

What sequences best represent environmental background?

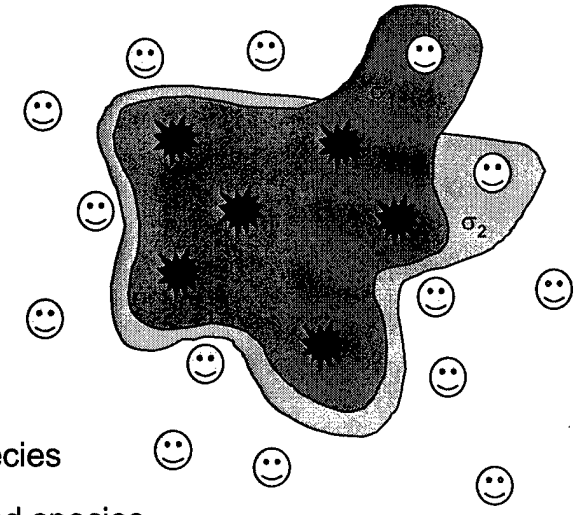
- Bacteria
- Viruses
- Insects
- Animal

Partition Target Sequences

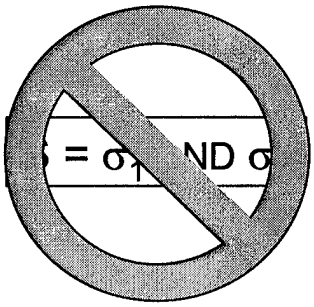


Isolate sequences unique to targets

Why search for unique sequence?



Non-unique signatures susceptible to false positives



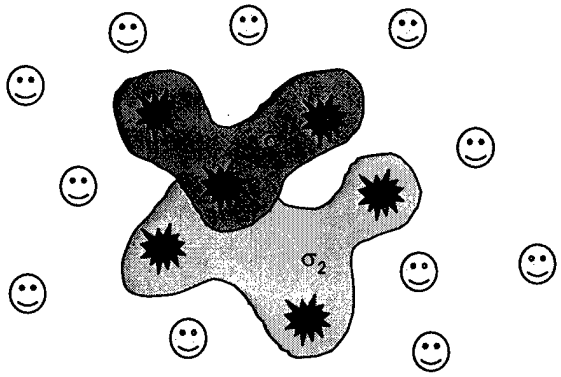
Target species

Background species

σ_1= ATGACCATGA ...

σ_2= GCATTCAGTA ...

Isolate sequences unique to targets



Detection Signature

$S = \sigma_1$ OR σ_2

Target species

Background species

σ_1= ATGACCATGA ...

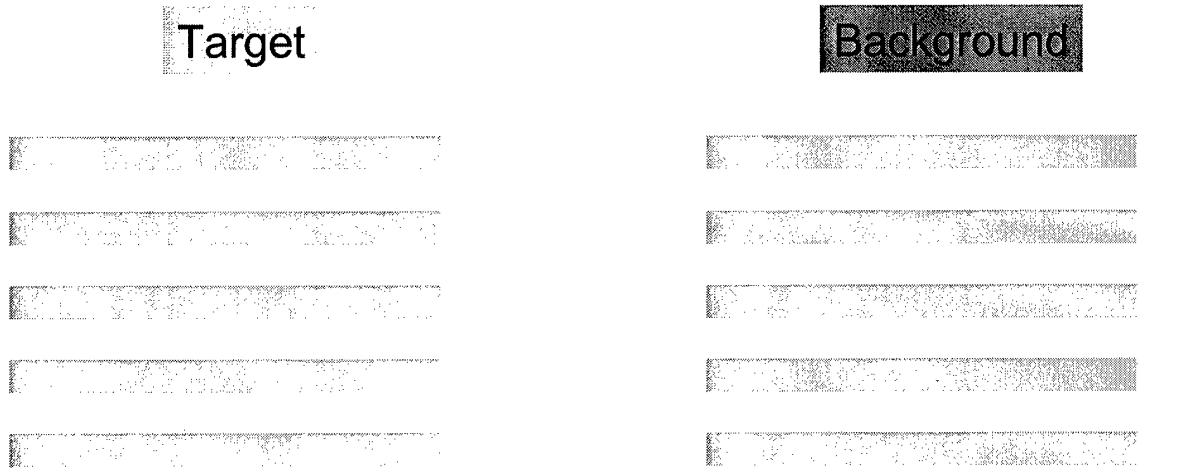
σ_2= GCATTCAGTA ...

Los Alamos
NATIONAL LABORATORY
EST. 1943

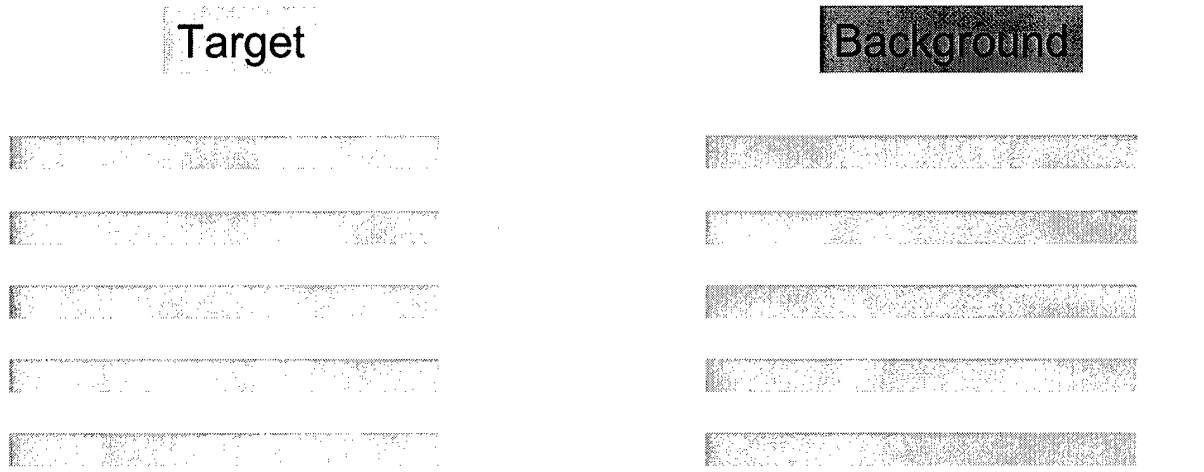
The World's Greatest Science Protecting America

NNSA

Isolate sequences unique to targets



Isolate sequences unique to targets



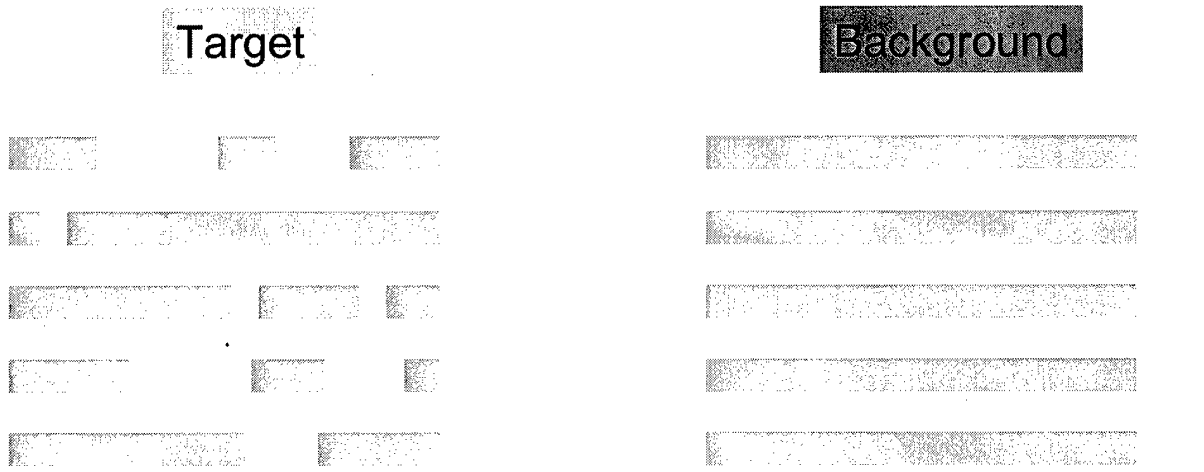
Los Alamos NATIONAL LABORATORY EST. 1943

Identify similar regions …

The World's Greatest Science Protecting America

NNSA

Isolate sequences unique to targets



and subtract background from target

Isolate sequences unique to targets

$$U = T - (T \; n \; B)$$

$\tilde{n}$?

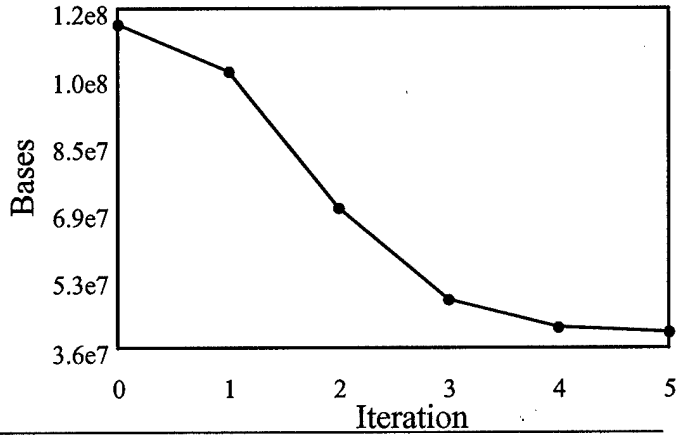
1. Use mpiBLAST to identify regions of similarity between U_i and B.
2. If a region has a T_m > 55 °C (over a = 21 base window) then it is considered an exact match.

$$U \sim U_f$$

$$U_{i+1} = U_i - (U_i \; \tilde{n} \; B)$$

$$U_0 = T$$

$$U_f \bullet U_3$$



Los Alamos NATIONAL LABORATORY EST 1943

The World's Greatest Science Protecting America

NNSA

Isolate sequences unique to targets

$$U = \bigcup_{i,j}^{N,\,M_i} f_{ij}$$

N = Number of target taxa
M_i = Number of contiguous sequence fragments, f, in species i

f_{0j} = {*E. coli* 042_0, *E. coli* 042_1, *E. coli* 042_2, …, *E. coli* 042_j, …}
f_{1j} = {*S. typhi*$_0$, *S. typhi*$_1$, *S. typhi*$_2$, …, *S. typhi*$_j$, …}
f_{2j} = {*Y. pestis* $CO92_0$, *Y. pestis* $CO92_1$, *Y. pestis* $CO92_2$, …, *Y. pestis* $CO92_j$, …}
⋮

Identify Signature Candidates

$$U = \bigcup_{i,j}^{N,\,M_i} f_{ij}$$

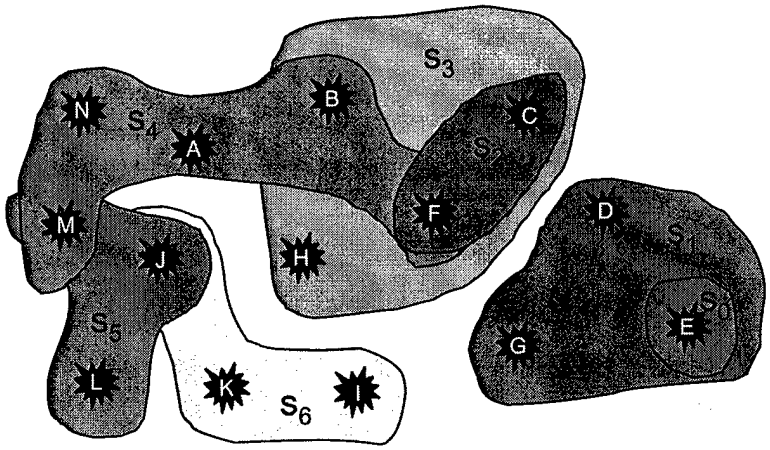
N = Number of target taxa
M_i = Number of contiguous sequence fragments, f, in species i

f_{0j} = {*E. coli* 042_0, *E. coli* 042_1, *E. coli* 042_2, ..., *E. coli* 042_j, ...}
f_{1j} = {*S. typhi*$_0$, *S. typhi*$_1$, *S. typhi*$_2$, ..., *S. typhi*$_j$, ...}
f_{2j} = {*Y. pestis* $CO92_0$, *Y. pestis* $CO92_1$, *Y. pestis* $CO92_2$, ..., *Y. pestis* $CO92_j$, ...}
⋮

Perform "all-against-all" multiple sequence alignments of f_{ij} against f_{kp}, i ? k

No thermodynamic criteria employed, 100% sequence identity required!

Find the minimal set of signatures that will detect all target taxa



s_i is a set of taxa that contain a common sequence fragment, σ_i

Find the minimal set of sequence fragments that will detect all target taxa

Given:

$T = \{A, B, C, D, E, F, G, \ldots, \}$
$s_0 = \{E\}$
$s_1 = \{D, E, G\}$
$s_2 = \{B, C, F, H\}$
$\vdots$

Our task is to find $G = \{s_\alpha, s_\beta, s_\delta, \ldots\}$

Where $T = \bigcup_i G_i$
and
$|G|$ is a global minimum

The set coverage problem is NP hard!

Find the minimal set of sequence fragments that will detect all target taxa

Approximate solution

Solve for G using Metropolis Monte-Carlo driven simulated annealing

Trial move generation: G_i ? G'_{i+1}

Add random ([0,n]) number of sets

and

Delete random ([0,n]) number of sets

Acceptance criteria: $G_{i+1} = G'_{i+1}$

$|G'_{i+1}| = |G_i|$

or

$\mathcal{R} = \exp(-[|G'_{i+1}| - |G_i|]/T)$

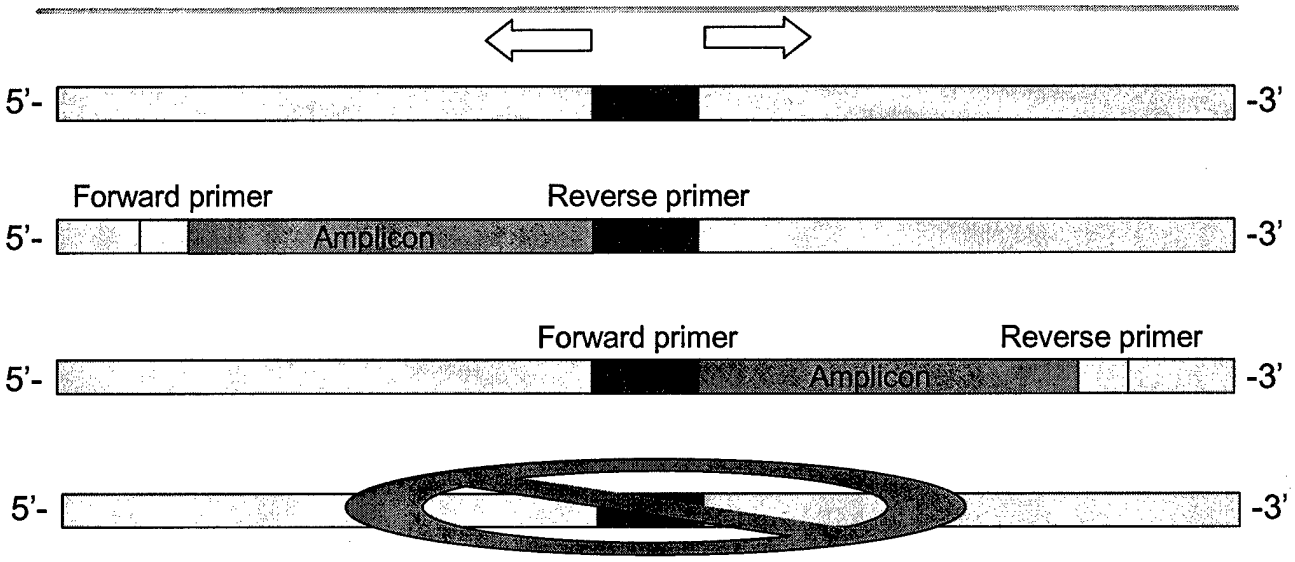
$\mathcal{R} \in [0,1]$

Validate computed signatures against all available DNA sequence data

1) Compute T_m for all signatures in G against all available DNA sequences (i.e. all of Genbank).
   - Identify false positives missed by greedy, BLAST-based comparison
   - Identify taxa that *should* be included in the background (but weren't)

2) Remove "promiscuous" signatures and re-compute G.

Filter signatures with experimental constraints: PCR Primers

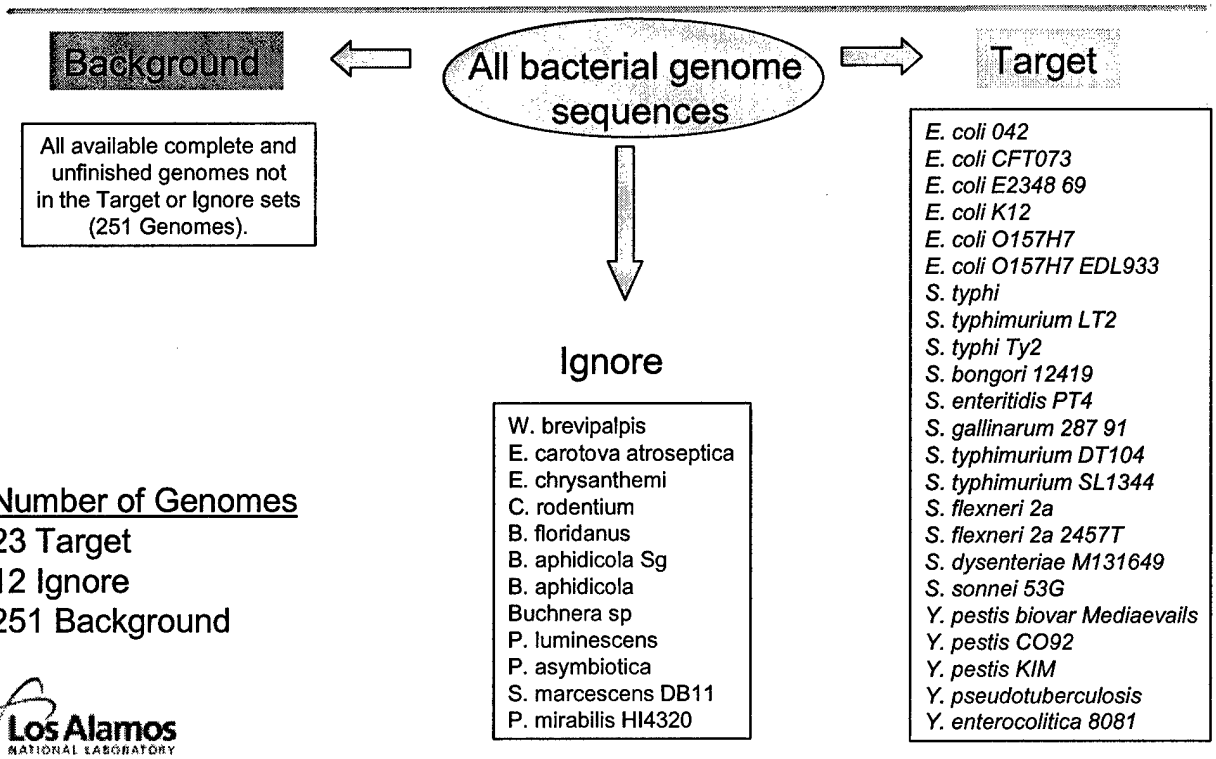


Los Alamos NATIONAL LABORATORY EST. 1943

The World's Greatest Science Protecting America

NNSA

Case study: enteric bacteria
Input genomes



Number of Genomes
23 Target
12 Ignore
251 Background

Los Alamos NATIONAL LABORATORY EST. 1943

The World's Greatest Science Protecting America

NNSA

Signature summary for enteric bacteria

125 signatures (unique regions with matching primers) in common for *all* enteric targets

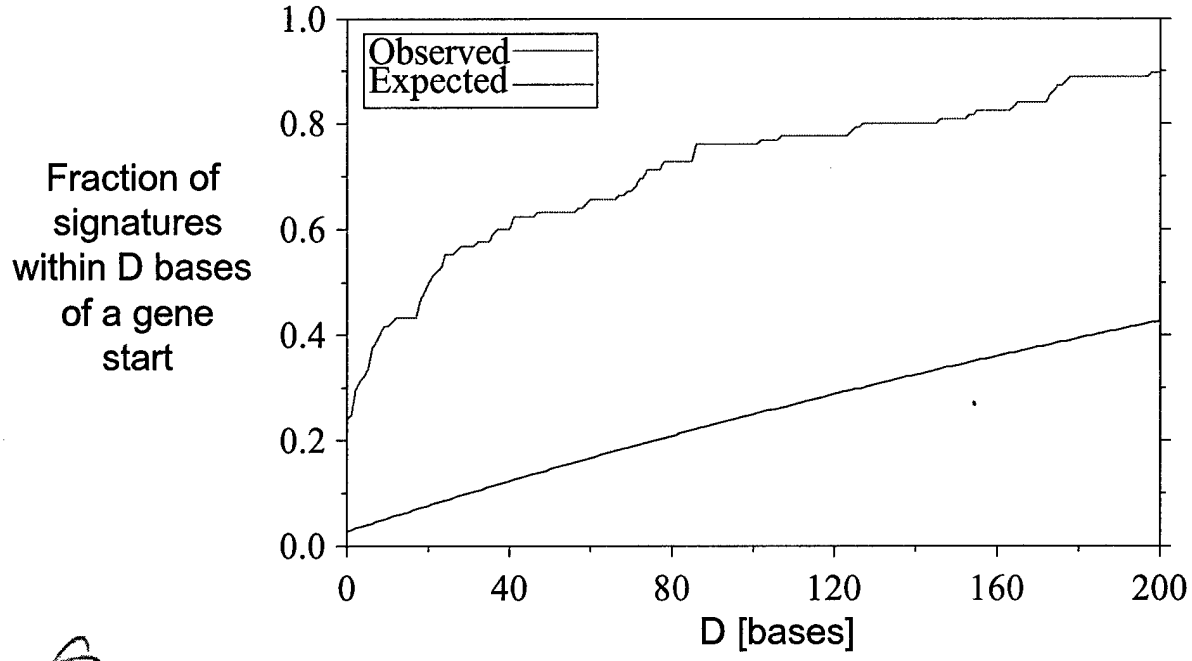
- 69 unique regions are contained in genes
- 23 unique regions are contained in intergenic space
- 33 unique regions span intergenic space and one or more genes

Signatures are *not* uniformly distributed

	Contained in gene(s)	Contained in intergenic space	Spans intergenic space/gene boundary
Expected*	**87%**	**10%**	**3%**
Observed	**55%**	**18%**	**26%**

*For *E. coli K12* and 21 base signature size

Signatures are preferentially located near a gene start

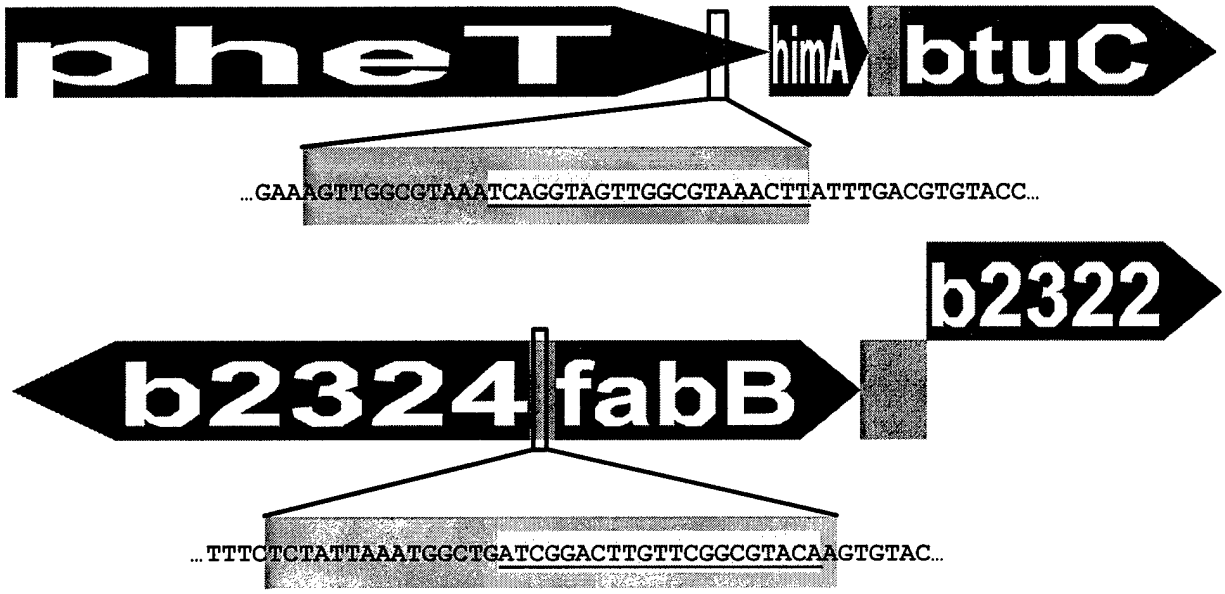


For *E. coli K12* and 21 base signature size

Enteric signature location bias: Transcription and translation elements

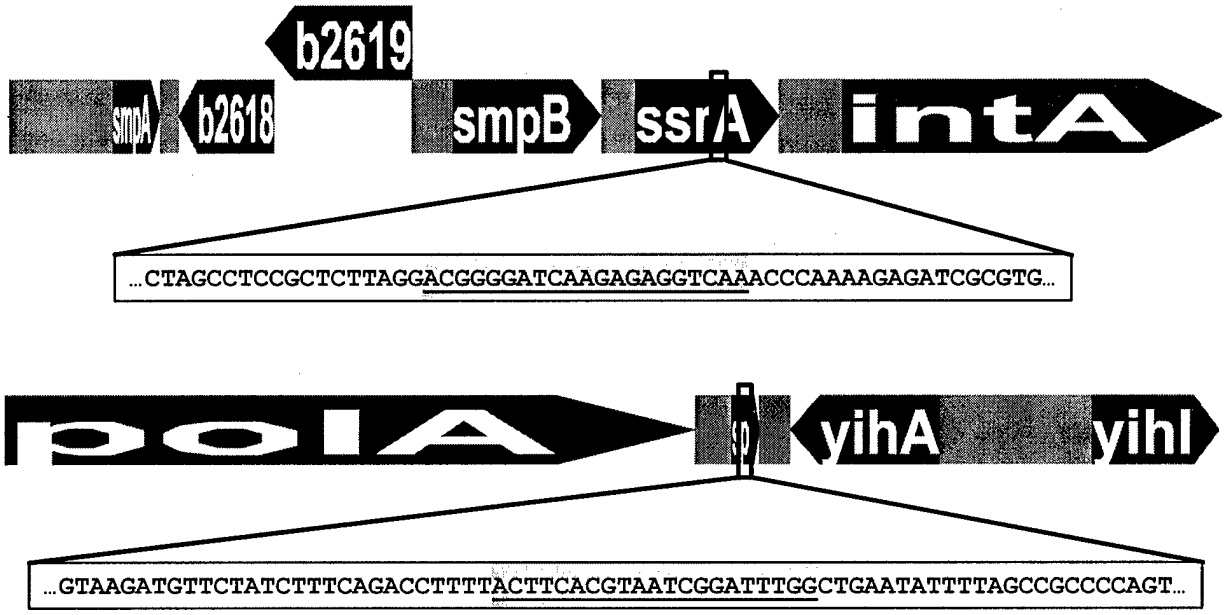
Signature "anchor"	Gene	Intergenic space	Gene/intergenic space boundary
Transcription factor binding site	X	X	
small RNA	X		
Ribosomal binding site		X	X

Transcription factor binding site
Example: himA and fabB



Los Alamos
NATIONAL LABORATORY
EST. 1943
The World's Greatest Science Protecting America
NNSA

small RNAs
Example: ssrA and spf

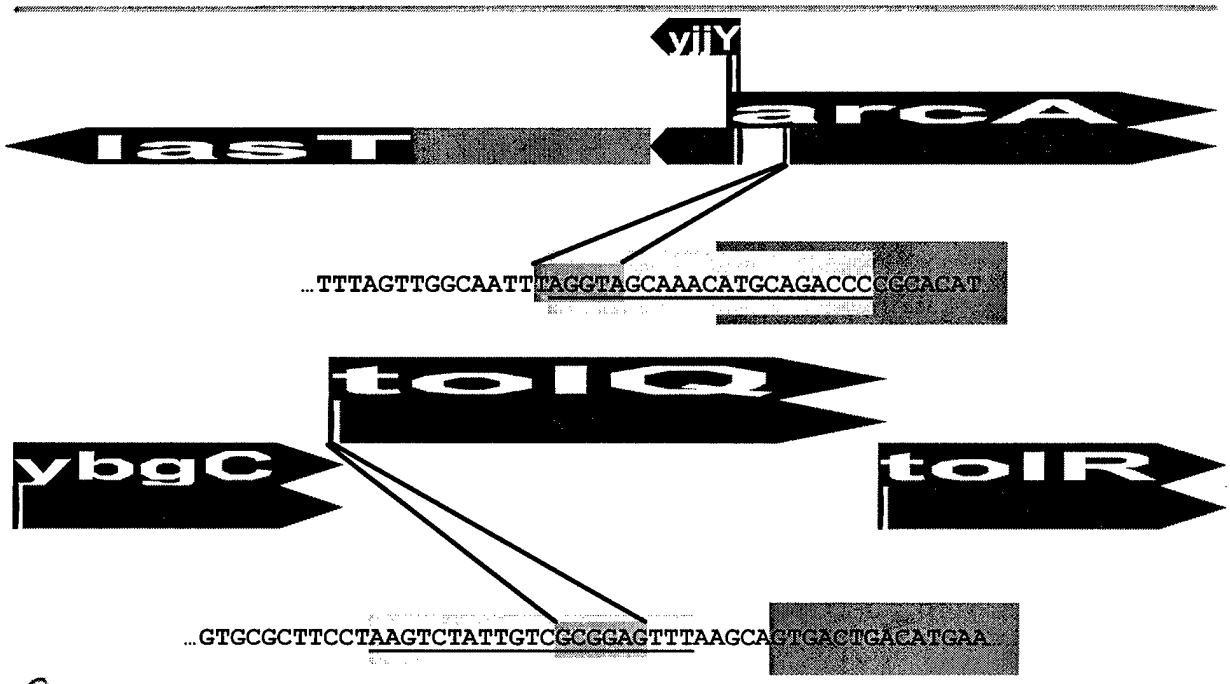


Los Alamos
NATIONAL LABORATORY
EST. 1943

The World's Greatest Science Protecting America

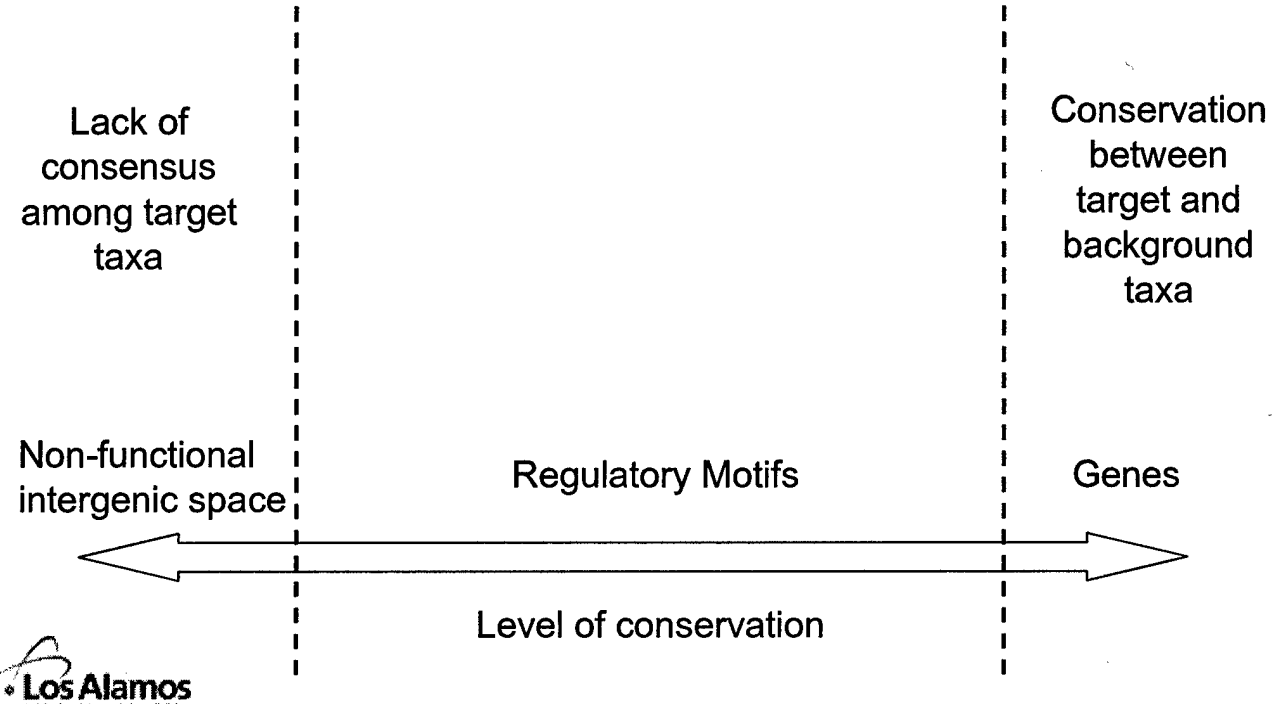
NNSA

Ribosomal binding sites
Example: RBS_{arcA} and RBS_{tolQ}

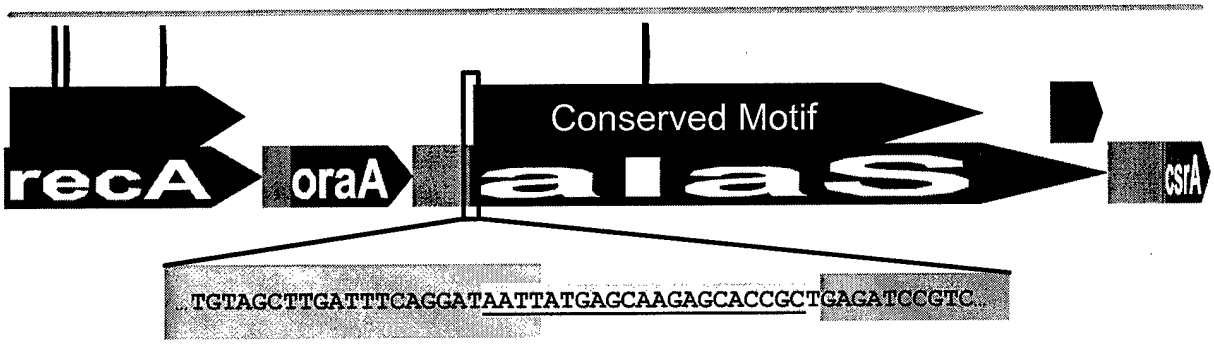


Los Alamos
NATIONAL LABORATORY
EST. 1943
The World's Greatest Science Protecting America
NNSA

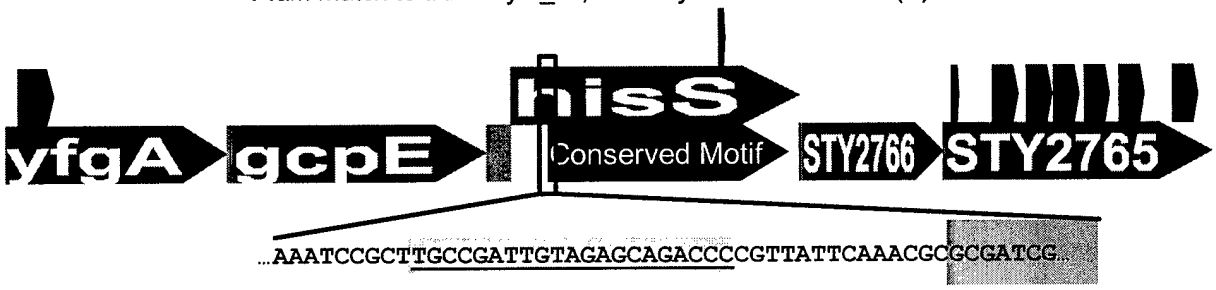
Why are signatures abundant at or near gene boundaries?



Gene boundaries
Example: alaS and hisS



Pfam match to tRNA-synt_2c, tRNA synthetases class II (A)



Pfam match to tRNA-synt_2b, tRNA synthetase class II (G, H, P, S and T)

Acknowledgements

Los Alamos National Laboratory
Bioscience Division

Informatics Team
Cathy Cleland
Norman Dogget
Rob Leach
Jian Song
Charlie Strauss
Chris Stubben
Murray Wolinsky*
Yan Xu
Wendy Zheng

Experimental Validation
John Dunbar*
Lance Green
Scott White

Funding
Department of Homeland Security
Department of Defense

The World's Greatest Science Protecting America

NNSA