

## Genome analysis

## Evolutionary analysis of enzymes using Chisel

Alexis A. Rodriguez<sup>1,2,\*</sup>, Tanuja Bompada<sup>1</sup>, Mustafa Syed<sup>1</sup>, Parantu K. Shah<sup>3</sup> and Natalia Maltsev<sup>1,2,\*</sup><sup>1</sup>Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, IL 60439,<sup>2</sup>Computation Institute, The University of Chicago, 5640 S. Ellis Avenue, RI 405, Chicago, IL 60637 and<sup>3</sup>Department of Human Genetics, The University of Chicago, 920 E. 58th Street, Chicago, IL 60637, USA

Received on May 25, 2007; revised on July 17, 2007; accepted on August 13, 2007

Advance Access publication September 13, 2007

Associate Editor: John Quackenbush

## ABSTRACT

**Motivation:** Availability of large volumes of genomic and enzymatic data for taxonomically and phenotypically diverse organisms allows for exploration of the adaptive mechanisms that led to diversification of enzymatic functions. We present Chisel, a computational framework and a pipeline for an automated, high-resolution analysis of evolutionary variations of enzymes. Chisel allows automatic as well as interactive identification, and characterization of enzymatic sequences. Such knowledge can be utilized for comparative genomics, microbial diagnostics, metabolic engineering, drug design and analysis of metagenomes.

**Results:** Chisel is a comprehensive resource that contains 8575 clusters and subsequent computational models specific for 939 distinct enzymatic functions and, when data is sufficient, their taxonomic variations. Application of Chisel to identification of enzymatic sequences in newly sequenced genomes, analysis of organism-specific metabolic networks, ‘binning’ of metagenomes and other biological problems are presented. We also provide a thorough analysis of Chisel performance with other similar resources and manual annotations on *Shewanella oneidensis* MR1 genome.

**Availability:** Chisel is available for interactive use at <http://compbio.mcs.anl.gov/CHISEL>. The website also provides a user manual, clusters and function-specific computational models.

**Contact:** arodri7@mcs.anl.gov or maltsev@mcs.anl.gov

**Supplementary information:** Additional data can be found at <http://compbio.mcs.anl.gov/CHISEL/htmls/refs.html>

## 1 INTRODUCTION

Evolutionary analysis of a wide spectrum of phylogenetically diverse organisms is essential for understanding adaptive strategies employed by organisms inhabiting different environments. Common ancestry of eukaryotes, prokaryotes and archaea led to similarity of many molecular functions. However, differences in organisms’ structural complexity, physiology and lifestyle have resulted in divergent evolution

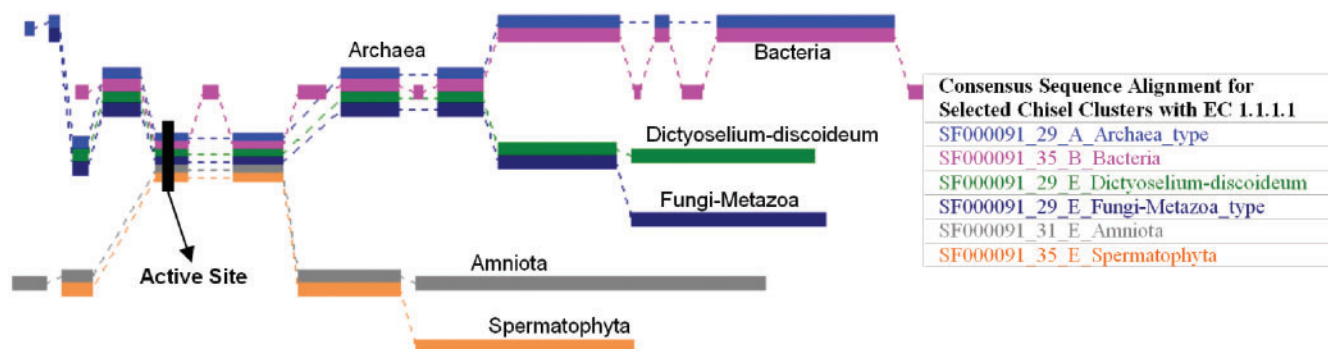
and emergence of variations of molecular function, metabolic organization and phenotypic features.

The availability of large volumes of sequence data has called for specialized methods for similarity-based annotation transfer. The simplest method for annotation transfer is the BLAST (Altschul *et al.*, 1997) or HMM-based similarity searches in primary sequence databases (Bork and Koonin, 1998). A number of pattern or motif databases, such as Prosite (Hulo *et al.*, 2006), PRINTS (Attwood *et al.*, 2003) and Blocks (Henikoff and Henikoff, 1996) are also available for function annotations. Several methods that utilize manual curation in sequence databases, evolutionary information, motifs and sophisticated algorithms have been proposed (Nariai *et al.*, 2007). For example, PRIAM (Claudel-Renard *et al.*, 2003) method uses annotations and domain information to perform clustering and provides PSSM matrix for each entry in ENZYME (Bairoch, 2000) database. EFICAz (Tian *et al.*, 2004) on the other hand focuses on recognition of functionally discriminating residues in enzyme families obtained by a conservation-controlled HMM iterative procedure for enzyme classification. However, primary focus of these and other methods is accurate inference and annotation of function.

On the other hand, availability of large volumes of sequence and enzymatic data for taxonomically and phenotypically diverse organisms should also allow study of emergence of function variation governed by environmental factors such as, temperature, mineral composition and more (Felsenstein, 1985). Such studies can provide insights into the adaptive mechanisms that led to the diversification of enzymes, in terms of their kinetic and enzymatic properties, subunit composition, cofactor preferences and other properties. Therefore, tools for high-resolution comparative and evolutionary analysis are required for the characterization of the molecular variations of enzymatic functions specific to taxonomic groups and phenotypes (Galperin and Koonin, 1999).

To this end, we present the Chisel system—an integrated bioinformatics environment and clustering pipeline for identifying and characterizing enzymatic sequences and their evolutionary variations, and subsequently the metabolic pathways. Analysis of enzymatic sequences in Chisel provides the basis for reasoning about the evolutionary history of an enzymatic

\*To whom correspondence should be addressed.



**Fig. 1.** The POAVIZ alignment (Grasso *et al.*, 2003; Lee *et al.*, 2002) of the consensus sequences for the Chisel clusters derived from the superfamily of zinc-containing alcohol dehydrogenases enzymes (PIR superfamily SF000091, EC 1.1.1.1). Each color represents a consensus sequence for a cluster of sequences corresponding to different taxonomic groups of organisms. The alignment demonstrates the conservation of the active site throughout the Chisel cluster consensus sequences and substantial variability in the N- and C-terminus of the sequences depending on their taxonomic origin.

function and answer questions such as ‘What variants of the same enzymatic functions have preferential use in certain phylogenetic neighborhoods or in a particular ecological niche?’ The Chisel system generates function- and taxonomy-specific clusters of enzymatic sequences and subsequent computational models. These models are presented to the user in a highly annotated and visualized form. Chisel supports both automated and interactive analysis of the data while providing tools for community curation of resulting models.

Chisel forms part of the PUMA2 (Maltsev *et al.*, 2006) integrated system for evolutionary analysis of metabolism and the development of organism-specific metabolic reconstructions. Exploration of the evolutionary history of enzymatic functions, in a larger context of metabolic pathways, is important for comprehending the evolution of particular metabolic processes and their variations. Identification of these variations provides insight into the emergence of differences in taxonomy- or phenotype-specific metabolic pathways. Such differences may then be exploited for microbial diagnostics, metabolic engineering and drug design. Another area which could greatly benefit from the described approach is in metagenome analysis. Identification of phenotypic and taxonomic variations of enzymes can improve methods for ‘binning’ of metagenomic data and assist in the development of descriptive models of metabolic networks characteristic for microbial communities (Tyson *et al.*, 2004; Venter *et al.*, 2004).

For example, three major evolutionary versions of alcohol dehydrogenases (ADH EC 1.1.1.1) are known: Zn-containing, Fe-containing and short-chain-type alcohol dehydrogenases. However, even evolutionarily close Zn-containing ADHs vary significantly. A Chisel alignment (Fig. 1) representing taxonomic-based analysis of Zn-containing alcohol dehydrogenases shows that while sharing conserved regions, including the active site location, sequences have undergone significant modifications. These function and taxonomy-specific sequences can be easily identified and grouped by Chisel (Fig. S1). For ADHs such variation among taxonomic groups may allow the organisms to adapt in different environments (see Supplementary Material).

The following sections describe our approach to evolutionary analysis of enzymatic sequences and metabolic pathways using Chisel. Examples of its applications to interpretation of genomes, metagenomes and microbial diagnostics are also presented.

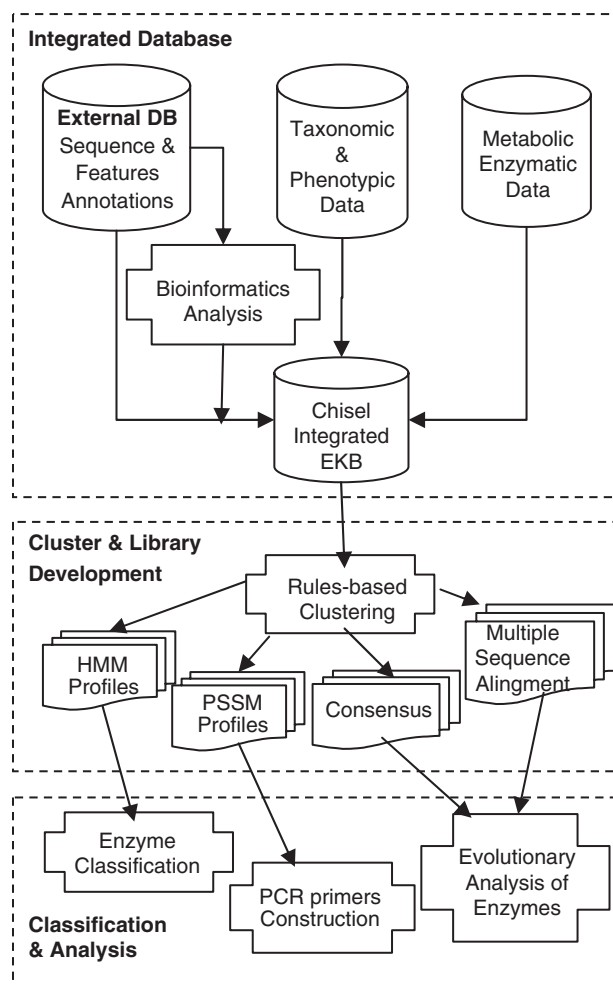
## 2 SYSTEM AND METHODS

Chisel is a Web-based bioinformatics system available at <http://compbio.mcs.anl.gov/CHISEL>. The system includes the following components (Fig. 2):

- An enzymatic knowledge base (EKB).
- A rules-based hierarchical clustering pipeline for identification of enzymatic functions and their taxonomic and phenotypic variations.
- A library of computational models for classification of un-annotated sequences.
- A web-based user interface with a suite of tools for interactive identification, comparative and evolutionary analysis, and annotation of the enzymatic sequences by expert users.

### 2.1 Description of the enzymatic knowledge base

The EKB is an integrated database of annotated enzymatic sequences that provides the data necessary for the rules-based clustering of enzymatic sequences and annotation of the resulting models. The EKB leverages the PUMA2 integrated database and contains following databases: (a) *metabolic and enzymatic information* from EMP (Selkov *et al.*, 1996, 1997), KEGG (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2000) and ENZYME databases and literature; (b) *sequence data and annotations* (e.g. functional domains, active sites, binding sites, experimental data) from NCBI (Pruitt *et al.*, 2005; Wheeler *et al.*, 2006), Gene Ontology (GO) (Midori *et al.*, 2000), PIR (Wu *et al.*, 2004, 2006) and UniProt (Apweiler *et al.*, 2004) databases; (c) *structural information* from PDB (Berman *et al.*, 2000), and structural classification from CATH (Pearl *et al.*, 2002) and SCOP (Murzin *et al.*, 1995); (d) *taxonomic information* and (e) *phenotypic information* (e.g. environmental niche, oxygen, pathogenicity, temperature and salinity requirements) from NCBI, TIGR and literature. In addition to the integrated information from these databases,



**Fig. 2.** Chisel system architecture includes (a) an integrated database of sequence data, annotations and enzymatic and metabolic information; (b) clustering pipeline and the libraries of resulting computational models and (c) a classification and analysis module containing tools for interactive analysis, refinement and annotation of the developed models.

the EKB also contains pre-computed results of sequence analysis by standard tools like BLAST, Blocks, InterPro (Mulder *et al.*, 2005), PSORT (Nakai and Horton, 1999) and TMHMM (Krogh *et al.*, 2001). The EKB data is accessible through the PUMA2 system. EKB is updated with major updates of the underlying databases and the update process is completely automated.

## 2.2 Chisel algorithm implementation

The Chisel rules-based pipeline performs high-resolution clustering of initial seed sets of homologous sequences into similarity-based clusters (see Figs S3–S7 for Chisel pseudocode and a detailed example of the clustering procedure). The resultant clusters are function specific and, when sufficient sequence data is available, they are function- and taxonomy-specific (i.e. contain sequences performing the same enzymatic function as annotated by the EC numbers and originating from organisms sharing the same taxonomic group). The computational models of enzyme functions are generated at the end of the clustering

procedure. The steps required in the development of the Chisel clusters are as follows:

### Step 1—Annotation of sequences from the initial set

The initial or seed sets of homologous sequences used by the Chisel clustering pipeline are annotated with information from the EKB. The sequences are also analyzed by an array of sequence analysis tools as mentioned in the previous section. The following classes of information from EKB/PUMA2 database are considered by the clustering pipeline as features:

- General information—sequence ID, sequence function (EC number, description), GO term, sequence length, organism of origin (taxonomy ID);
- global similarity—iProClass superfamily, relevant COG ID(s), top 10 BLAST hits;
- sequence features—length, InterPro domains (domain ID, location), Blocks results (blocks ID, location, Z score), feature type (i.e. binding site, active site) and location. Domain definitions from PSORT and membrane regions predicted by TMHMM.

### Step 2—Clustering of the enzymatic sequences

The procedure applied for the clustering of seed sets of enzymes according to their function and taxonomic origin is a hierarchical pipeline (see Figs S4–S6). The clustering procedure includes the following:

**2.2.1 Initial clustering of seed sets** Homologous sets of sequences are clustered based on the composition and location of the domains in the sequences. The clusters with uniform domain composition (i.e. same order of domains present) are used for the next level of clustering. This procedure enables significant increases in the quality of the multiple sequence alignments (MSA) in further steps of clustering (Fig. S3). In order to keep uniformity, only sequences with <15% difference in length are included in the alignments. Large difference in length of sequences with similar domain composition may be due to its structural or phenotypic changes and such sequences are treated separately.

**2.2.2 Function and taxonomy-based clustering** The initial clusters may contain multiple functions. In order to achieve function-based clustering, the sequences belonging to each initial cluster are aligned using ClustalW (Thompson *et al.*, 1994) or MUSCLE (Edgar, 2004). The neighbor-joining dendrogram (Saitou and Nei, 1987) is partitioned by extracting branches containing sequences with a common enzymatic function (annotated with identical EC number). Additionally, the function-specific branches containing sequences originating from the same taxonomic group (kingdom and lower) are also extracted. In order to achieve maximal separation between the resulting Chisel clusters we utilized two distance measures: separability and compactness to measure the quality of resultant clusters. Ancestral protein sequences generated using ANCESCON were used to derive both the distance measures (Cai *et al.*, 2004). The phylogenetic distance between the (super) family ancestral sequence and ancestral sequences for each cluster is taken as the measure of separability. Distances between each sequence in the cluster and cluster ancestral sequence (Cai *et al.*, 2004) is measured as cluster compactness. The clusters for which the compactness measurement is less than the separability measurement are kept.

During the clustering process, sequence outliers with distinctive composition of domains or sequences showing low parsimony values are separated. This outlier sequences are useful for further biological investigation as they may reveal unique properties of a particular

enzyme or might be incorrectly annotated. Thus, the process of classification can also help in identifying errors in annotations.

### 2.3 Development of Chisel models and the classification of unannotated sequences

The function- and taxonomy-specific clusters resultant of the clustering process is used to derive Chisel models of enzyme function. Chisel models for a particular enzymatic function may contain multiple taxonomy-specific clusters. All Chisel clusters are annotated with a unique identifier, a multiple sequence alignment (MSA) using ClustalW. They also include MSA-based hidden Markov model (HMM) profiles generated by HMMER (Eddy, 1996, 1998); a position-specific scoring matrices (PSSM) (Gribskov *et al.*, 1987); Blocks profiles (Henikoff and Henikoff, 1996) and consensus and ancestral sequences (Cai *et al.*, 2004; Henikoff and Henikoff, 1996) at a desired sequence identity cutoff representing the clusters.

The initial sets of homologous sequences for the development of Chisel models could be obtained from public resources or users can provide them in an interactive session. The current sets of Chisel clusters are derived from the PIR iProClass enzymatic protein superfamilies (release 2.82). These clusters were generated from a seed set of over 2.5 million sequences originating from 1930 PIR superfamilies. These seed sets often contain sequences performing different functions and originating from a variety of taxonomic groups. In present version, Chisel resolves these superfamilies into 8575 clusters and subsequent computational models specific for 939 distinct enzymatic functions and, when data is sufficient, for their taxonomic and phenotypic variations. The sizes of the clusters vary from 4 to 150 sequences.

The clusters can be used in automated or interactive analysis. These models allow users to perform different types of analysis including sequence annotations, evolutionary analysis and design of oligonucleotide primers. For example, Chisel computational models (e.g. HMM profiles, PSSMs) can be used to classify unannotated sequences through the Chisel web-based user interface. The Blocks profiles generated by Chisel provide a basis for the development of oligonucleotide primers using the CODEHOP program (Henikoff and Henikoff, 1996) to support experimental research.

The clustering of sequences through the Chisel pipeline and the development of computational models is a CPU-intensive task. In order to reduce the time required for generation of accurate models, the GADU (Genome Analysis and Database Update) server (Sulakhe *et al.*, 2005) was utilized. GADU provides a gateway to the Grid to perform the computationally intensive tasks required by Chisel. The use of Grids allows immediate access to the required computational cycles on an opportunistic basis. For example, the clustering of sequences and subsequent model building in the latest release of Chisel took 14h on 100 CPUs. Classification analysis of sequences of an average prokaryotic genome using HMMER and the Chisel library of HMM profiles can take 2h on 100 CPUs.

Automated update of Chisel clusters and computational models will be carried out with updates on the EKB. Subsequent releases of Chisel will allow the development of models from other libraries of protein families including COGs (Tatusov *et al.*, 1997) and Hobagen (Perrière *et al.*, 2000). Inclusion of new sequences to the clusters strengthens the models and in some cases, finds new enzymatic models which were non-existent in due to the lack of sequences with similar features.

### 2.4 Chisel user interface and workbench for interactive sequence analysis

The Chisel user interface offers a variety of tools for navigation and interactive data analysis of enzymatic sequences. In an interactive Chisel session, users can specify seed sequences, clustering features like

gene ontology, COG IDs, top blast hits, phenotypes, binding and active sites. We are working towards automatic specification of these additional features in the clustering process. While, the procedure for rule-based clustering as well as for the generation of models is automated, experts can subsequently refine both the steps using additional tools available at the Chisel website.

Besides searches based on taxonomy, keywords and sequences, an advanced search allows navigation of the Chisel clusters based on the physiological features (e.g. environmental niche, oxygen requirements, temperature and salinity preferences) (see Supplementary Material). Such a view of the data may be useful for identifying variations of enzymatic functions associated for particular phenotypes (e.g. thermostable enzymes, enzymes associated with pathogenesis). Projection of the Chisel taxonomy-specific clusters onto the metabolic networks available via 'Explore Pathways' may contribute to studies of evolution of taxonomy-specific metabolic pathways. Clusters in Chisel are extensively annotated with metabolic, phenotypic and enzymatic data. Sequence features are displayed to the users in a graphical form.

Chisel also offers unique tools (PhyloBlocks, BlocksBlast, Dragonfly) that allow the users to develop, refine and annotate the models interactively. These tools enable interactive assessment of the quality of Chisel models and developing models from user-submitted sets. Such corroboration and updates of the models by the experts are critical for the development of high-quality models. We plan to establish direct connections with scientific authorities in the fields of genetic sequence analysis and enzymology to enable quality control and validation of the developed models. Chisel also contains tools for classification of unannotated sequences based on the libraries of Chisel HMM models. These tools are accessible from the Chisel web-based interface and allow both analyses of individual sequences and batch submissions of sequences for annotation.

## 3 BENCHMARKING

The function- and taxonomy-specific clusters of enzymatic sequences obtained in the described process were used as a training set for the development of the Chisel models. The performance of the Chisel algorithm was then validated in many different ways. First, we compared the annotations provided by Chisel clusters to manual annotations with the jackknife approach (Zhang and Chou, 1995). Each sequence was tested against each cluster generated per experiment to test if the correct function was assigned to the sequence. Then, we compared taxonomy and functional specificities of Chisel models to several enzymatic protein families and domain libraries. Finally, performance of Chisel was also benchmarked on all enzymatic sequences of *Shewanella oneidensis* MR1 genome for which manual annotations are available.

Protein sequences that have their functions experimentally verified constitute the best and most reliable training set. There fore, PIR superfamilies containing at least two sequences with experimentally established functions were selected for the jackknife approach. Both the learning and test subsets were assured to have at least one sequence with experimentally verified protein function. Testing was performed with a total of 19905 experimentally verified protein sequences (annotated with experimental GO evidence codes and extracted from references in the BRENDA database). These sequences were resampled during the jackknife analysis a number of times, depending on the size of the PIR superfamily, to achieve

**Table 1.** Functional specificity of enzymatic protein families and domain libraries

EC/family	1 EC	2 EC	3 EC	≥4 EC
InterPro (5436 families)	<0.01% (20)	38% (2065)	19% (1051)	43% (2300)
Pfam (2828 families)	<0.01% (6)	40% (1134)	19% (532)	41% (1156)
PIRSF500000 (151 families)	1% (2)	52% (77)	21% (32)	26% (40)
PRIAM (3019 families)	100.0% (3019)	0% (0)	0% (0)	0% (0)
Chisel (8575 families)	98.4% (8438)	1.4% (120)	0.2% (17)	0% (0)

accuracy in testing experiments and to generate a larger sample of sequences. The experiment was repeated 201 950 times.

In the context of these experiments, a correct function assignment constitutes a match in enzyme nomenclature number with the experimentally verified annotation (i.e. true positive); a true negative constitutes each time an experimentally verified enzyme was not classified as a non-matching enzyme function. Functions were predicted correctly for 94.28% of the samples. The experiments resulted in a sensitivity measure of 95.8% and a specificity measure of 99.1% (see Supplementary Material). The false negatives were due in large part to the insufficient number of sequences for the development of Chisel models for a particular enzymatic function or its taxonomic variation in the learning period. The false positives were sequences predicted with an incorrect function or taxonomic group. Such false positive results may be explained by the lack of a model for a 'correct' function as a result of an insufficient training set for its development, causing false positive prediction of 'next to correct' function in cases of evolutionarily related enzymes. We plan to explore a number of approaches to overcome such over predictions. One approach is to augment the resolution of Chisel by increasing the number of sequence features to be considered by Chisel's pipeline.

We have performed comparisons of Chisel clusters with a number of protein family resources that generate their families by automated methods, such as PIR iProClass (Wu *et al.*, 2004), along with commonly used domain libraries (e.g. InterPro, Blocks). These domain libraries have proved to be extremely useful for automation of genetic sequence and evolutionary analysis of proteins. The quantity of enzymatic functions associated with individual protein families from InterPro (release 14.0), Pfam (release 21.0), PRIAM (release July 2006), and Chisel is presented in Table 1. The PIR subfamilies (release 2.82) containing protein clusters within a homeomorphic family (Wu *et al.*, 2004) having specialized functions and/or variable domain architectures (PIRSF ≤500000) were also included in the comparison. Only families of enzymatic sequences were used in the comparisons.

Table 1 shows in parentheses the number of enzymatic functions associated with each protein family developed by various groups. For example 20 InterPro families were specific for one enzymatic function (EC). As it follows from a table Chisel clusters have a very high degree of functional specificity in comparison to the other systems investigated: 98.4% of Chisel clusters are function specific. The PRIAM

clusters are generated from individual entries from ENZYME database corresponding to identical function and therefore provide 100% specificity. However, due to the low number of sequences available in the ENZYME database, many of the clusters contained single sequences, which may reduce the sensitivity. The Chisel clusters associated with more than one enzymatic function contain multifunctional enzymes. Our analysis has demonstrated that a significant percentage of the protein families from the investigated resources contain sequences associated with two or more enzymatic functions.

In addition, we compared the taxonomic specificity of protein families developed by the above-mentioned groups. The lowest common taxonomic node for the sequences in the protein families was reported. For consistency, we have taken into consideration only three taxonomic levels: the root or cellular organism, kingdom and subkingdom levels. Results from this experiment show that Chisel has a significantly higher resolution in identification of taxonomic variations of enzymes. Most of the Chisel clusters correspond to taxonomy lower than the kingdom taxonomic levels. For example, the superfamily PIRSF500093 (ATP synthase beta chain) contains sequences with a lowest common taxonomic level of 'cellular organism'. The Chisel pipeline recognized substantial differences within this superfamily and split it into clusters belonging to *Proteobacteria*, *Alphaproteobacteria*, *Gammaproteobacteria*, *Burkholderiales*, *Firmicutes*, *Bacillaceae*, *Cyanobacteria*, *Spermatophyta*, *Viridiplantae* *Magnoliophyta*, *Bangiophyceae* and *Bilateria*. Additional material for this comparison can be seen in the Supplementary Material. Performance of Chisel was also benchmarked on all enzymatic sequences of *S.oneidensis* MR1 genome for which manual annotations are available. There was 90.7% agreement between Chisel annotations and manual annotations. More over, when manual annotations were considered as a golden standard, Chisel achieved Specificity of 90.29% and Sensitivity of 84.55% (see Supplementary Material).

#### 4 RESULTS AND DISCUSSION

Identification of taxonomic and phenotypic variations of enzymes has already proved useful for a number of applications. Examples of such applications will be presented in the following sections.

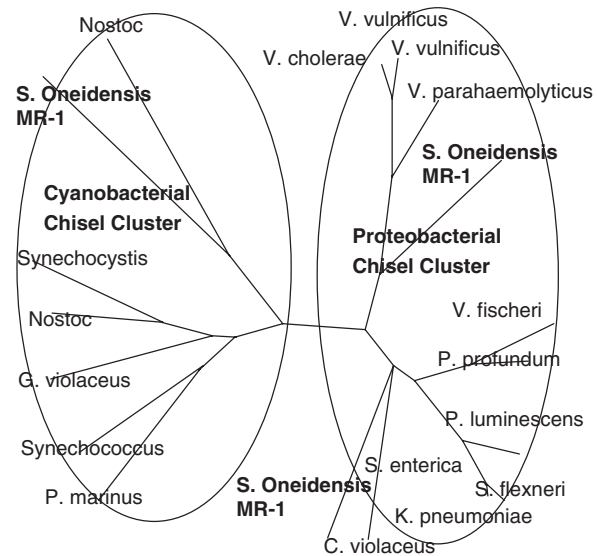
## 4.1 Analysis of genomes

Identification of taxonomic and phenotypic variations of enzymes has already proved useful for a number of applications. The following examples show areas in which the Chisel system has provided an added benefit to their analysis.

**4.1.1 Prediction of gene functions** Chisel was used as a supplemental tool for automated annotation of 12 *Shewanella* genomes in GNARE (Sulakhe *et al.*, 2005) for the *Shewanella* Federation. *Shewanella* is a *Gammaproteobacterium* that can grow both aerobically and anaerobically utilizing a diversity of electron acceptors (nitrite, nitrate, thiosulfate, iron, manganese, uranium). The study of these organisms presents a unique opportunity to investigate the adaptation of these metabolically versatile organisms to the environment. Analysis of *Shewanella* genomes using Chisel has identified 8476 enzymatic sequences (out of 43 691 sequences total) in 12 genomes. The number of predictions for individual genomes ranged from 718 in *S. denitrificans* OS217 to 823 in *S. oneidensis* MR1.

**4.1.2 Identification of taxonomy-specific metabolic signatures** Out of 823 enzymes predicted by Chisel in *S. oneidensis* MR1, 273 proteins corresponded to clusters specific for *Proteobacteria* and 129 proteins to *Gammaproteobacteria*. Such variations in levels of taxonomic specificity indicate that enzymes in the metabolic pathways evolve at different rates. In the course of adaptation, some of the enzymes become more specific for particular taxonomies. In *Shewanella*, the *Gammaproteobacteria*-specific variations of enzymes are associated predominantly with core metabolic pathways (e.g. glycolysis, purine and pyrimidine biosynthesis, biosynthesis of amino acids), as well as chemotaxis and sensory transduction processes. Chisel analysis of glycolytic pathways demonstrated that *S. oneidensis* MR1 contained the *Gammaproteobacteria*-specific versions of glycolytic enzymes phosphoglycerate mutase (EC 5.4.2.1) and acetyl-transferring pyruvate dehydrogenase (EC 1.2.4.1), while other enzymes of glycolysis were similar to other *Proteobacteria* [e.g. pyruvate kinase (EC 2.7.1.40), fructose-bisphosphate aldolase (EC 4.1.2.13), glucose-6-phosphate isomerase (EC 5.3.1.9)]. *S. oneidensis* contained two versions of alcohol dehydrogenase (EC 1.1.1.1): proteobacterial version of iron-containing ADH and *Gammaproteobacteria*-specific bifunctional aldehyde/alcohol dehydrogenase (see Supplementary Material for the complete list of taxonomy-specific clusters in *Shewanella*). This observation suggests significant systems-level adaptation that led to diversification of enzymes in this group of organisms in the course of evolution. Identification of taxonomy-specific signature enzymes may provide insights into mechanisms driving the emergence of taxonomy and phenotype-specific pathways. These signatures may also be used for microbial diagnostics and metabolic engineering.

**4.1.3 Discovery of potential cases of horizontal gene transfer** The Chisel analysis also helps to identify potential cases of horizontal gene transfer. The Chisel clustering pipeline allows inclusion of up to 10% of sequences from organisms different from the predominant taxonomy of the sequences in



**Fig. 3.** Phylogenetic tree for peptide deformylase sequences (EC 3.5.1.88). *Shewanella* peptide deformylase were classified by the Chisel pipeline in two clusters: a proteobacterial cluster (SF004749\_6\_B\_Proteobacteria8) containing two versions of the enzyme from *S. oneidensis* and a cyanobacterial cluster (SF004749\_6\_B\_Cyanobacteria4) composed of cyanobacterial sequences and one *S. oneidensis* sequence. The following sequences were used for constructing the phylogenetic tree *Nostoc* sp. PCC 7120, *Synechocystis* sp. PCC 6803, *Gloeobacter violaceus*, *Prochlorococcus marinus*, *Synechococcus* sp. WH 8102, *S. oneidensis* MR-1, *Photobacterium profundum*, *Vibrio fischeri*, *Shigella flexneri*, *Klebsiella pneumoniae*, *Photobacterium luminescens*, *Chromobacterium violaceum*, *V. cholerae*, *V. vulnificus*, *V. parahaemolyticus* and *Salmonella enterica*.

the cluster. These sequences represent candidates for horizontal gene transfer. For example, our analysis of the *S. oneidensis* genome identified three sequences of peptide deformylase (EC 3.5.1.88). This prokaryotic enzyme removes the formyl group from the N-terminal Met of newly synthesized proteins. Two of the three sequences performing this function belonged to the proteobacterial Chisel cluster (SF004749\_6\_B\_Proteobacteria8), while another version of the same enzyme belonged to the Cyanobacterial version of the cluster (SF004749\_6\_B\_Cyanobacteria4), suggesting that this enzyme might have been acquired via horizontal gene transfer. The distribution of the sequences from these clusters on the phylogenetic tree is represented in Figure 3. We plan to systematically evaluate Chisel predictions in order to identify potential cases of horizontal gene transfer in other organisms.

**4.1.4 Identification of enzymatic subunits and isozymes** Chisel proved to be useful in identification of enzymatic subunits and isozymes. For example, Chisel identified 11 sequences of ATP synthases from *Mus musculus* and classified them in seven ATP synthase clusters: alpha chain, mitochondrial precursor; beta chain, mitochondrial precursor; gamma chain, mitochondrial precursor; lipid-binding protein subunit C, mitochondrial precursor; vacuolar subunit A; vacuolar subunit D and vacuolar subunit F.

## 4.2 Interpretation of metagenomes

A major problem in metagenome analysis is ‘binning’, i.e. assigning correct taxonomic origin to the genomic fragments. This process is usually performed by analysis of 16S RNA sequences (if available) or phylogenetic analysis of BLAST hits.

Chisel presents an additional method of analysis of metagenomes by predicting taxonomic variations of enzymes in the sample.

Chisel was used for the analysis of metagenomes from contaminated sediments beneath a leaking high-level radioactive waste tank at the DOE Hanford site [<http://compbio.mcs.anl.gov/PNNL1>]. Because of the extremely harsh environmental conditions (i.e. high radiation, salinity and temperature levels) the amount of biomass levels and the number of sequences available for analysis were low. Therefore, thorough analysis of available data was especially important in order to predict taxonomic distribution and physiological properties of organisms residing in such extreme environments.

The Chisel analysis of a high-contamination zone metagenome identified 543 enzymatic sequences corresponding to 263 distinct enzymatic functions. The predominant taxonomic groups of organisms identified in the sample were *Actinomycetales* (28%, corresponding to 152 Chisel predictions) and *Bacillus* (22%, corresponding to 122 Chisel predictions). Other predicted groups included a number of hits from extremophilic organisms: *Deinococcus* (6 hits), *Euryarchaeota* (16 hits) and *Symbiobacterium thermophilum* (6 hits). The predicted enzymes in the most abundant groups (*Actinomycetales* and *Bacillus*) corresponded to the core metabolic pathways. These results match the results predicted by 16S RNA analysis of this data. Chisel allows for further investigation of this metagenome by supporting the design of taxonomy-specific oligonucleotides to be used by our collaborators at PNNL in experimental component of this project. These degenerative oligonucleotides are based on the alignments of sequences corresponding to taxonomy-specific Chisel clusters.

## 4.3 Identification of taxonomy-specific variations of enzymes for biomedical research

Identification of the variations of enzymatic functions of pathogenic organisms (e.g. *Enterobacteriaceae*, *Staphylococci*) is important for many biomedical applications, including microbial diagnostics and recognition of potential antibacterial drug targets. Chisel has a substantial number of clusters containing sequences originating from pathogenic organisms. For example, the current release of Chisel contains 247 models for enzymes specific for *Enterobacteriaceae* and lower taxonomic groups (e.g. 11 clusters for *Salmonella*, 9 for *Escherichia coli*), 90 models for *Staphylococcus* and 126 models for *Streptococcus* groups of organisms. A significant number of the functions represented by these models correspond to functions essential for the livelihood of these organisms. The pool of enzymes in these clusters represents a set of potential targets for antibacterial therapies. The Chisel system supports the development of PCR primers and oligonucleotides

corresponding to these models using the CODEHOP program (Henikoff and Henikoff, 1996).

This feature can assist experimentalists in identifying pathogenic organisms using biochip- or PCR-based technologies. We also note that the vast majority of microorganisms cannot be grown as a pure culture, thus making the study of their physiology and metabolism difficult. The development of a library of alignments, sequence profiles and data for development of libraries of oligonucleotides for taxonomic variations of enzymes may assist in characterizing organisms whose identities have not been established yet or those that cannot be cultured in the laboratory.

## 5 CONCLUSIONS

To our knowledge Chisel is the first specific resource for identifying taxonomic and phenotypic variations of enzymes.

The Chisel system has the following important features distinguishing it from other systems:

- Representation and analysis of Chisel models in the framework of metabolic pathways give a systems-level perspective on the evolution of metabolic pathways, enzymes and related protein families.
- The developed libraries allow for the construction of degenerative PCR primers. These primers can be used to support *in vitro* bacteriological diagnostics and characterization of microorganisms.
- Chisel supports community curation of the models using interactive tools (e.g. PhyloBlocks).

Chisel performs well on different benchmark experiments and we suggest several applications to it. We are developing the system further to provide fine-grained functional analysis and also utilizing it for pathway analysis across genomes.

## 6 AVAILABILITY

The Chisel system is available for interactive use at <http://compbio.mcs.anl.gov/CHISEL>. The Chisel clusters are updated regularly with each update from the Uniprot/PIRSF database. The libraries of Chisel models (HMM, PSSM, consensus sequences, multiple sequence alignments) are available for download at the Chisel website.

## ACKNOWLEDGEMENTS

We are very grateful to Prof. Wen Hsiung Li (The University of Chicago) for his invaluable advice and suggestions in the course of the project. We extend special thanks to the following individuals who contributed valuable advice and support: Elizabeth Glass, Mark D'Souza, Dinanath Sulakhe and Yi Zhang. We also acknowledge Luke Ulrich for his contribution in the development of PhyloBlocks and BlocksBlast. We are grateful to Dr M. Galperin for his thoughtful comments and advice in the preparation of this manuscript. We are very grateful to Prof. Wen Hsiung Li (The University of Chicago) for his invaluable advice and suggestions in the course of the project. This work was funded in part by the Chicago

Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust; by the US Department of Energy under Contract DE-AC02-06CH11357 and by the National Science Foundation under grants 86044 (GriPhyN), 122557 (iVDGL) and the NCSA Alliance Expedition 'A PACI Petascale Data Quest' (PDQ).

*Conflict of Interest:* none declared.

## REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Apweiler,R. *et al.* (2004) UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
- Attwood,T.K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
- Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
- Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences – where are the bottlenecks? *Nat. Genet.*, **18**, 313–318.
- Cai,W. *et al.* (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol. Biol.*, **4**, 33.
- Claudel-Renard,C. *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.
- Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
- Eddy,S.R. (1998) Profile Hidden Markov Models. *Bioinformatics*, **14**, 755–763.
- Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
- Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.
- Galperin,M.Y. and Koonin,E.V. (1999) Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica*, **106**, 159–170.
- Grasso,C. *et al.* (2003) POAVIZ: a partial order multiple sequence alignment visualizer. *Bioinformatics*, **19**, 1446–1448.
- Gribskov,M. *et al.* (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
- Henikoff,J.G. and Henikoff,S. (1996) Blocks database and its applications. *Meth. Enzymol.*, **26**, 88–105.
- Hulo,N. *et al.* (2006) The PROSITE database. *Nucleic Acids Res.*, **34**, D227–D230.
- Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
- Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
- Lee,C. *et al.* (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–64.
- Maltsev,N. *et al.* (2006) PUMA2 – grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res.*, **34**, D369–D372.
- Midori,H. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- Mulder,N.J. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
- Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Nakai,K. and Horton,P. (1999) PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–35.
- Nariai,N. *et al.* (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, **2**, e337.
- Pearl,F.M. *et al.* (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Sci.*, **11**, 233–244.
- Perriere,G. *et al.* (2000) HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.*, **10**, 379–385.
- Pruitt,K.D. *et al.* (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
- Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Schomburg,I. *et al.* (2000) Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Funct. Dis.*, **3–4**, 109–118.
- Selkov,E. *et al.* (1996) The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Res.*, **24**, 26–28.
- Selkov,E. *et al.* (1997) The metabolic pathway collection: an update. *Nucleic Acids Res.*, **25**, 37–38.
- Sulakhe,D. *et al.* (2005) Gnare: automated system for high-throughput genome analysis with grid computation backend. *J. Clin. Monit. Comput.*, **19**, 361–369.
- Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
- Thompson,J.D. *et al.* (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Tian,W. *et al.* (2004) EFICAZ: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.*, **32**, 6226–6239.
- Tyson,G.W. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
- Venter,J.C. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
- Wheeler,D.L. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
- Wu,C.H. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
- Wu,C.H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
- Zhang,C.T. and Chou,K.C. (1995) An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. *J. Protein Chem.*, **14**, 583–589.