Supplementary Material

Rodriguez *et al.*, Bioinformatics

1. Benchmarking of the Chisel system and comparisons to other functional clusters

2. Supplementary Discussion

1. Benchmarking of the Chisel system and comparisons to other functional clusters

Benchmarking

Several methods can be used to evaluate the effectiveness of Chisel's predictions for unannotated sequences (e.g., ROC analysis (Egan, 1975), jackknife and bootstrap (Zhang and Chou, 1995)). The ROC method requires the availability of large amounts of experimentally verified data. In our case, the experimentally verified data constitutes only 10% of the total data. Furthermore, the experimentally verified data is highly biased toward a few eukaryotic organisms (i.e., *H. sapiens, A. thaliana*, *M. musculus, R. norvegicus)* and model prokaryotic organisms (i.e., *E. coli* and *B. subtilis*).

Chisel clusters were validated by using the jackknife approach (Zhang and Chou, 1995). The iProClass PIR superfamilies (Wu *et al*., 2004) containing at least two sequences with an experimentally established function were selected. The jackknife experiments were performed on all sequences from these superfamilies used for the development of Chisel models. Both the learning and test subsets were assured to have at least one sequence with experimentally verified protein function. This strategy gives a higher confidence value to the evaluation process.

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. Protein sequences that have their functions experimentally verified constitute the best and most reliable training set. Testing was performed with a total of 19,905 experimentally verified protein sequences (annotated with experimental GO evidence codes (Midori *et al*., 2000) and extracted from references in the BRENDA database (Schomburg *et al*., 2000)). These sequences were resampled during the jackknife analysis a number of times, depending on the size of the PIR superfamily, to achieve accuracy in testing experiments and to generate a larger sample of sequences.

True Positives	190,397
True Negatives	375,869
False Positives	3,124
False Negatives	8,429

The experiment was repeated 201,950 times. Each sequence was tested against each cluster generated per experiment to test if the correct function was assigned to the sequence. 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 experiment resulted in 190,397 true positives; 375,869 true negatives; 8,429 false negatives; and 3,124 false positives. This gave a sensitivity measure of 95.8% and a specificity measure of 99.1%. 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 overpredictions. One approach is to augment the resolution of Chisel by increasing the number of sequence features to be considered by Chisel’s algorithm (e.g., existence of transmembrane domains, additional feature location).

4.1 Comparison of Chisel to Similar Resources

To our knowledge no other resource is being developed specifically for identifying taxonomic and phenotypic variations of enzymes. We have performed comparisons of Chisel clusters with a number of protein family resources, such as PIR iProClass (Wu *et al.*, 2004), TIGRFAM (Haft *et al.*, 2003), along with commonly used domain libraries (e.g., InterPro, BLOCKS (Henikoff and Henikoff, 1996)), which 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 12.1), PFAM (release 21.0), TIGRFAM (release 6.0), PRIAM (release July 2006) (Claudel-Renard *et al.*, 2003), 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 annotated enzymatic functions (ECs) in the protein families developed by each group. Chisel clusters have a significantly higher degree of functional specificity (i.e., a single enzymatic function associated with the cluster) in comparison to the other systems investigated, with 98.4%

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)
TigrFam (1511 families)	1% (12)	44% (663)	26% (394)	29% (442)
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)

of clusters being function specific. 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 enzymatic functions. The PIRSF homeomorphic families were found to be the most specific among the compared resources.

In addition, we compared the taxonomic specificity of protein families developed by the above-mentioned groups. The lowest common node for the member 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. The results of these comparisons are presented in Table 2. The table shows that Chisel has a significantly higher resolution in identification of taxonomic variations of enzymes. Most of the Chisel clusters correspond to lower than kingdom taxonomic levels. For example, the superfamily PIRSF500093 (ATP synthase beta chain) contains sequences with a lowest common taxonomic level "cellular organism." The Chisel algorithm recognized substantial differences within this superfamily and split it into clusters belonging to Proteobacteria, Alphaproteobacteria, Gammaproteobacteria, Burkholderiales, Firmicutes, Bacillaceae, Cyanobacteria, Spermatophyta, Viridplantae Magnoliophyta, Bangiophyceae, and Bilateria.

Table 2. Taxonomic specificity of enzymatic protein families from InterPro, PFAM, TIGRFAM, PIRSF subfamilies, and Chisel.

Common Tax Level	**Root or Cellular Organism**	**Kingdom**	**Sub-kingdom**
InterPro (5436 families)	77.5% (4217)	12.9% (703)	9.6% (516)
Pfam (2828 families)	75.6% (2140)	14.4% (408)	10% (280)
TigrFam (1511 families)	75.5% (1145)	16.8% (255)	7.7% (111)
PIRSF (151 families)	37% (56)	27.9% (42)	35.1% (53)
Chisel (8575 families)	0.07% (6)	16.6% (1421)	83.4% (7148)

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

(a) Representation and analysis of Chisel models in the framework of metabolic pathways give a systems-level perspective on the evolution of enzymes and their protein families.

(b) The developed libraries allow for the construction of degenerative PCR primers. The primers support *in vitro* bacteriological diagnostics and characterization of microorganisms.

(c) Chisel supports community curation of the models using interactive tools (e.g., PhyloBlocks).

2. Supplementary Discussion

The abundance of genomic and enzymatic data has led to the development of several algorithms that performs grouping or clustering of related sequences for a variety of applications including homology identification, study evolutionary conservation, homology modeling, and domain detection etc. for function prediction (Narai et al., 2007). The Chisel rules-based pipeline performs high-resolution clustering of initial seed sets of homologous sequences into similarity-based clusters. Clustering algorithms are generally applied to group functionally related sequences and providing sensitive profiles for further utilization. Several general purpose tools, such as BLAST, PSI-BLAST (Altschul *et al*., 1997) and HMM-based profile are routinely used to group/identify similar sequences with sequence libraries like TIGR and PFAM. They are traditionally used for sequence similarity based domain detection. Additional tools have been developed to tackle the issue of sequence function annotation. However, methods that quantify similarity by using some attribute of the best BLAST hit and use single-linkage clustering. Systers and GeneNest provide clusters reflecting overall sequence similarity and not necessarily the function. Depending on the method and thresholds used the clustering results may vary. Such straightforward approaches can also group together different multi-domain proteins which share a common domain, and can be fooled by promiscuous domains (Doolittle, 1995; Marcotte *et al*., 1999). To this end, Chisel uses domains detected by the Interpro tool in its first step to avoid this pit-fall.

Existing tools such as PRIAM take the clustering a bit further and provide a PSSM based library to annotate the sequence by function. PRIAM forms sub-clusters of enzymes based on the domain composition of the enzymes coming from the ENZYME database (Bairoch *et al.*, 2000). Thus, all of their clusters should be function-specific. Similar to Chisel, one main clustering feature of PRIAM is the domain composition. However, unlike PRIAM, Chisel aims to answer additional questions including more nuanced functional classification, biological niche and environmental factors affecting the enzyme that allow it to behave similar to other taxonomically related organisms. Even though PRIAM contains many function-specific clusters (Table 1), Chisel provides far greater number of clusters and function specificity because more sequences are used for the analysis and taxonomic specificity is also sought.

Chisel is a very specific and sensitive tool to provide function and taxonomic classification of a given enzyme sequence. However, in the present version it doesn't identify functionally discriminative residues. Such tool like EFICAz (Tian *et al.*, 2004) that use clustering and subsequent HMM based iterative procedures to recognize functionally discriminating residues are currently available. However, their main focus is accurate inference of function annotations. Other applications of Chisel are discussed in the main manuscript.

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25(17), 3389-402.

Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, 28, 304-5.

Claudel-Renard, C., *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*, 31(22), 6633-9.

Doolittle, R.F. (1995). The multiplicity of domains in proteins. *Annu. Rev. Biochem.* 64: 287–314.

Egan, J.P. (1975) Signal Detection Theory and ROC Analysis, Academic Press, New York.

Haft, D.H. *et al.* (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, 31(1), 371-3.

Henikoff, J.G. and Henikoff, S. (1996) Blocks database and its applications. *Meth. Enzymology*, 26, 88-105.

Krause A, *et al*., (2002) SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res.* 30(1):299-300.

Marcotte, E.M., *et al.* (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science* 285: 751–753.

Nariai, N., *et al.* (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE* (2) e337.

Midori, H., *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genet.* 25, 25-9.

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, 109-18.

Tian, W, *et al*. (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res.* 32(21) 6226-39.

Wu, C.H. *et al*. (2004) PIRSF: Family classification system at the Protein Information Resource. *Nucleic Acids Res.*, 32, D112-4.

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, 583–9.