Using high performance computing and domain-based functional annotation of proteins to enhance discovery of novel proteins, identify functional homology, and characterize phylogenetic relatedness

TR-10-02
Jeffrey L Tilson, Gloria Rendon,
Eric Jakobsson
June 8, 2010

RENCI Technical Report Series
http://www.renci.org/techreports

Jeffrey L Tilson, Renaissance Computing Institute,
University of North Carolina at Chapel Hill

Gloria Rendon, National Center for Supercomputing
Applications, Urbana-Champaign, U.S.A.

Eric Jakobsson, NCSA, Beckman Institute for
Advanced Science and Technology,
Department of Molecular and Integrative Physiology,
University of Illinois at Urbana-Champaign, U.S.A.

Using high performance computing and domain-based functional annotation of proteins to enhance discovery of novel proteins, identify functional homology, and characterize phylogenetic relatedness

Jeffrey L Tilson[1]*, Gloria Rendon [2]*, Eric Jakobsson[2,3]*§

[1] Renaissance Computing Institute, University of North Carolina at Chapel Hill, U.S.A.

[2] National Center for Supercomputing Applications, Urbana-Champaign, U.S.A.

[3] Beckman Institute for Advanced Science and Technology, Department of Molecular and Integrative Physiology, University of Illinois at Urbana-Champaign, U.S.A.

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

JLT: jtilson@renci.org

GR: rendong@ncsa.uiuc.edu

EJ: jake@ncsa.uiuc.edu

Abstract

Background

Next generation sequencing technology is putting significant pressure on computational researchers to implement software tools for analysis (identification, annotation, homology/orthology assignment, phylogeny, etc.) of the genes and gene products “on-the-fly” in parallel with the sequencing machines. This requires both leveraging supercomputing systems and alternative kinds of analyses. We seek to

contribute to the solution of these problems through the deployment of high speed explicitly functional domain-based solutions through the system called MotifNetwork. We present case select studies of domain-based approaches to gene analysis that range from homology assessment to phylogeny reconstruction to pangenomic analysis as a demonstration of potential benefits of such approaches. For analyses, we used grid-computing to enable the computations necessary to apply these techniques to genome-size systems.

Results

We used MotifNetwork to apply functional domain-based methods to three biological test cases that represent broad biological areas of research.

- First, we assess functional homology of over 3000 eukaryotic proteins with respect to the ligand-gated ion channel family by calculating domain-based similarity of genes with four different metrics: distinct-partners, inverse document coefficients, cumulative association coefficients, and the Jaccard function.
- Second, we illustrate a methodology for predicting phylogenetic relatedness based on evolutionary domain analysis. It is applied to over 40 prokaryotic proteins that were identified as likely functional homologs with respect to the same family of ion channels.
- Lastly, comparative genomics studies are conducted between. *H. sapiens* and 23 different strains of *E. coli*. The domain-based pangenome of *E. coli* is analyzed and compared against that of *H. sapiens* in a context of drug target identification and potential side effects.

Benchmarks of MotifNetwork indicate that execution times achieve reasonable performance scaling when using up to 256 processors available to this work and that

our use of a data-grid for storage of the results, as implemented with iRODS, is well-suited for large-scale biological pipelines.

Conclusions

The combination of domain-based analyses and fast processing enabled by MotifNetwork should permit researchers to more accurately and efficiently perform research on a wide range of biological problems and thus alleviate the bottlenecks that now exist between sequencing of genes and their subsequent characterization. Our approach is especially suitable for biological problems that can be formulated as the identification of functional correspondences among a large set of proteins such as the three illustrative examples that are discussed in the paper which range from *E. coli* pangenomics, to functional homology and phylogenetic relatedness of the LIC family of ion channels.

Background

The concept of domains within genes and proteins (conserved segments of the macromolecular sequences with a particular function and structural motif) has become increasingly important in sequence analysis in recent years. This perspective as applied to structural domains is reviewed in [1]. Most present-day proteins have multiple functional domains [2] and it appears that much of the evolutionary path of proteins occurs by rearrangement and addition (or, less usually, subtraction) of domains [3]. In fact, Fong et al [3] model protein evolution by a maximum parsimony analysis of domains contained in the protein, the results of which suggest that the overall path of evolution has been towards greater complexity (Fusions, which cause an increase in the number of domains in proteins, have outnumbered fissions that reduce the number of domains). Björklund et al [4] suggest the concept of “domain

distance" (degree of difference of a domain architecture) as a measure of evolutionary distance between proteins. Song et al. [5] and Lin et al. [6] also propose metrics for domain architectures in order to determine homology between proteins. A review of the literature, data, and concepts of domain recombination is provided in [7].

Our laboratory, in collaboration with the laboratory of L. Aravind, used domain-based methods to find prokaryotic members of the Ach receptor channel family after Blast-based techniques, utilizing complete eukaryotic protein sequences as probes, found nothing of interest. In particular, when [8] parsed the eukaryotic sequences into conserved domains, and did pattern matching utilizing the various domains, there was revealed a group of prokaryotic members of the acetylcholine receptor channel family. Subsequent experimental work confirmed the identity of this group [9]; Later an X-ray crystal structure was determined [10].

Part of the usefulness of these methodologies is tied to the speed of generating the results. For our larger research-scale genome studies, results could takes months to complete on our typical computers rather than the 1-2 days desired. In order to apply larger amounts of computing capacity to domain-based methods, we have developed the MotifNetwork system [11-12]. It consists of a suite of scientific workflows and custom applications using grid-computing concepts. A major component of MotifNetwork's biological applications suite is the InterProScan software [13-14] which is used to identify functional domains and motifs in protein sequences. MotifNetwork, however, is designed to use alternative definitions of domains such as *de novo* approaches using MEME [15] and MAST [16]. Workflows as referred to in this document are software instantiations of conceptual pipelines that perform a series of biologically relevant computations. The workflow orchestrator controls the entire computation and schedules all operations. Grid components assist by performing

basic operations such as the scheduling of compute jobs onto remote computers, the transfer of data between computers, and the assembly and storage of the results. The simplicity of these biological pipelines masks much of the underlying software complexity which must be uniformly speedy to ensure high performance of the system's workflows.

Results

The following three case studies illustrate the main theme of this paper for many biological problems that can be reformulated as determining functional relatedness of sets of proteins. Select aspects of the design and performance of MotifNetwork that impact the day-to-day usability of the system are also discussed.

Case study 1: Assessing functional homology using domain-based metrics

This analysis begins by assembling over 3000 eukaryotic proteins that share at least one functional domain with representatives of the ligand-gated ion channel (LIC) family, as defined by the TCDB Transporter Classification Database [17] (date of access, October, 2009). See Figure 1. Four different metrics were used for assessing functional homology; these consisted of two weighted metrics: distinct-partners and inverse document coefficient, and two unweighted metrics: cumulative association coefficient and Jaccard function. See Song et al. [5] and Lin et al. [6] for details on these metrics. We also run benchmarks on the metrics to assess their sensitivity and specificity.

All metrics identified distant homologs of the LIC family. The weighted metrics outperformed the unweighted metrics in sensitivity and specificity for cases where distinct domain set profiles existed to correctly assign homology to a subfamily of the LIC family. For example, they performed well with homologs of the ACH subfamily which has the distinct domain set [IPR002394 IPR006029 IPR006201

IPR006202 IPR018000], but less well with distinguishing homologs of some subfamilies with the same domain set profile. For instance, the GABA receptors in the 1.A.9.5 subfamily, the 5-HT-gated chloride channels, and the 1.A.9.6 subfamily, all have the same domain profile [IPR006028 IPR006029 IPR006201 IPR006202 IPR018000]; thus a protein with this domain set could belong to either one of these subfamilies. Groupings at a higher resolution than pure domain profiles can be obtained by first aligning sequences using domain profiles followed by cluster analysis on those alignments. Our experiences so far indicate that sequences with the same domain profiles invariably align readily

Case Study 2: Predicting likely prokaryotic homologs of the LIC family

To date, only two prokaryotic proteins have been experimentally verified to belong to the LIC family; they are GLVI (GenBank:NP_927142.1) and ELIC (PDB:2VL0). We collected all prokaryotic proteins that shared at least one domain with the LIC family, 1.A.9 as per TCDB [**16**] (date of access March, 2009) and after applying domain-based homology metrics were able to identify 42 proteins that are likely homologs of the LIC family, including GLVI or ELIC. The set of domains that comprise the composition for these likely homologs consists of only four domains [IPR006201, IPR006202, IPR006028, IPR006029]. See Figure 2. In comparing corresponding protein sequences from eukaryotes and prokaryotes, we typically find that the prokaryotic sequences have fewer domains and a simpler domain structure.

Case Study 3: Domain-based comparative genomics of the *E. coli* pangenome

At the time of the calculations presented in this section, 23 genomes of different strains of *E. coli* were completed and publicly available. They consisted of 15 pathogenic strains, four commensal strains, and four genomes often explored in the laboratory but that do not normally inhabit humans. In addition to the *E. coli*

genomes, we also considered that of *H. sapiens*. All genomes were processed by MotifNetwork to resolve the proteins into their functional domain components and then to calculate the universe of domains for the entire *E. coli* pangenome and for *H. sapiens*. Being able to identify *E. coli* proteins that do not contain human functional domains supports comparative studies across the species such as identifying drug targets with a lessened likelihood of side effects in humans. A summary of the domain content of the *E. coli* and *H. sapiens* genomes is shown in Figure 3. One noteworthy finding is that the human universe of functional domains is only approximately 50% larger than that of the *E. coli* pangenome, despite the number of identified human proteins being approximately five times larger than for a typical *E. coli* genome. The larger repertoire of human proteins appears to come from more combinations of domains in a single protein rather than from a larger set of domains. We have previously observed this relationship between eukaryotic and corresponding prokaryotic proteins in a different context [18]. The true ratio of *H. sapiens* to *E. coli* proteins is probably much greater than indicated in Figures 3-4 because of frequent splice variants produced during human transcription that rarely occur in bacteria.

One potential use of domain analysis is to identify potential targets for specific antimicrobial drugs as an alternative to the nonspecific antibiotics that are inducing antibiotic resistance in a variety of pathogenic microbes. Figure 4 is a Venn diagram that shows the commonalities and the differences between the domain content of human proteins, pathogenic and commensal *E. coli*. In this diagram, we see that there are 1602 domains that are common to all three; i.e., humans and both categories of *E coli*. About 25% of the domains in human proteins are shared with *E. coli*. For purposes of designing antimicrobial drugs, one interesting group of domains is the

326 domains that are in pathogenic but not in commensal *E. coli* or in humans. It is a reasonable presumption that knocking down or knocking out proteins containing those domains could influence the pathogenicity of those *E. coli* strains. A second group of interest is comprised of the 84 domains that are common to pathogenic *E. coli* and humans, but are not present in the commensal bacteria. Within these domains are those that subvert the host cell's signalling processes by mimicking host protein domains [19].

In our view, the fact that complexity in the human proteome arises from new ways to combine a relatively small number of domains, plus the fact that domain analysis seems able to pinpoint pathogenicity factors in bacteria, underscores the importance of studying and classifying proteins on the basis of domain composition and architecture.

MotifNetwork performance

Figure 5 reports the execution time for a significant step of one MotifNetwork workflow. Here a set of input proteins is scanned to identify, and record the location along the complete protein sequence, InterPro domains. The scanning step can be segmented into many smaller steps called *jobs*. The set of jobs defined in this way is considered an *ensemble*. The *width* of an ensemble is a measure of the number of independent jobs (subtasks) that may be simultaneously launched. The degree of runtime concurrency, however, can be less than this width by decreasing the number of available computing devices. An example of the scaling as a function of this constrained width is collected in Figure 5. All results were obtained on a cluster of Dell PowerEdge 1955 blades (2.66 GHz) using an InfiniBand interconnection network (PCI-Express SDR).

Here is processed the *E. coli* (4,226 non-duplicate input sequences) genome yielding 3,824 proteins that contain an Interpro integrated domain. A total of 3,861 domains were identified. The overall time to solution decreases substantially with the degree of concurrency. It shows that the execution time (runtime) for fairly small genomes including: scanning, data assembly, and post-processing steps, decreases from more than one day to approximately 3.5 hrs. It is important to note that for these results each node is actually a 4 processor system effectively increasing the performance and concurrency further still. Also reported is the relative efficiency. This is a measure of the effectiveness of increasing the number of available nodes on the runtime. A decrease to approximately 55% is observed. The efficiency decreases for two reasons. The first is overhead to manage the ensemble. Second, the ensemble width and the number of available compute nodes can poorly match causing load imbalances at the end of the run.

Figures 6 and 7 display the execution times for two applications that process results from the ensemble step. These are parallel programs based on the MPI standard [20]. The *ScoreMatrix* app constructs large data sets corresponding to the domain likelihood scores and positions versus the protein. The *WebMatrix* app constructs several graphs and properties of the identified domains. These include protein-domain and domain-domain co-location data and frequencies. Many of these datasets are in Cytoscape [21] compatible formats.

A common measure of parallel performance is speedup (SU). In particular, we want to characterize the effectiveness of increasing the number of applied processers to a

given calculation. Both figures report the measured wall time (sec) and speedup (SU) versus processors. Inefficiencies and load-imbalances decrease SU.

$$SU = \frac{t(p_o) \times p_o}{t(p)}$$

t(p)=walltime on p cores,

p= number of applied cores.

po= minimum number of cores

All results were based on the ensemble results for the *H. sapiens* genome. Excluding non-integrated InterPro domains results in 29,892 proteins that contain 3,566 unique domains. ScoreMatrix results are collected in Figure 5 indicating the walltime decreased from 32,000 sec to nearly 4,000 sec. The associated SU is computed to be nearly 250 at 256 applied processors. The WebMatrix results are displayed in Figure 7. Total wall time decreases from nearly 1,800 sec to 150 sec with a reported SU of nearly 210 at 256 processors. More extensive performance data are available in a recent report [22] and references therein.

Methods

Functional homology of the LIC family using domain-based analysis

We began with the set of nearly 30 proteins in the 1.A.9 family of TCDB to create a representative set of proteins of the ligand-gated ion channel (LIC) family. Each protein was resolved into its corresponding set of functional domains as defined by InterPro. A set of 19 InterPro domains is sufficient to represent all proteins in the LIC family in terms of their domain composition (see Table 1).

All proteins in the UniProtKB database (release 15.9) that contain at least one domain from the list in Table 1 were collected from the InterPro website and run with MotifNetwork. We further subdivided the proteins into two datasets: eukaryotic proteins (3104 proteins), and prokaryotic proteins (46 proteins). These were separately analyzed. To the eukaryotic proteins, we only applied four functional

homology metrics. To the prokaryotic proteins we further applied the evolutionary domain analysis.

Of the 3104 eukaryotic proteins, all but 470 are homologs of the LIC family. The set of 470 sequences all come from the probe 1.A.5.2 and contain the signature domain IPR004241 (Light chain 3), characteristic of a group of proteins that are sometimes associated with LIC's but share no homology with the LIC's. See Figure 1 for the results.

The set of prokaryotic proteins consisted of 46 sequences with forty of them likely homologs of the LIC family. They contain in their domain compositions at least one neurotransmitter ligand binding domain from [IPR006201, IPR006202] and zero or more neurotransmitter transmembrane domains from [IPR006028, IPR006029]. The remaining six proteins contained the neurotransmitter binding site conserved domain IPR018000 combined with a domain that is characteristic of restriction modification enzymes: IPR014883. Identification of these six illustrates the strengths of domain search techniques; It uncovers novel combinations of domains whose existence would not readily be inferred by whole sequence alignment with Blast.

Domain-based comparative genomics of *E. coli*

We processed each of the 23 *E. coli.* proteomes with MotifNetwork. In addition, we included *H. sapiens* for comparison. We proceeded to calculate the domain-based pangenome of *E. coli*. We also calculated separately the domain-based pangenomes of the entire set of *E. coli* strains, and further separately the pathogenic strains, the lab-use-only and the commensal strains of *E. coli*. Lastly, we identified domains unique to

each genome. The results are shown in Figures 3-4. The set of domains unique to the pathogenic strains provides potential targets for drugs that can potentially attack pathogenic *E. coli* without side effects on either the human host or on commensal *E. Coli*

MotifNetwork

The *WholeGenome* workflow of MotifNetwork was used to process the *E. coli* datasets and, the *ProteinProbe* workflow to run the LIC homologs. These workflows were created and are orchestrated using the Taverna workbench [23], their grid-architecture [24] is built on Globus grid-services [25] using GT4 [26], and supercomputing resources were allocated from the Renaissance Computing Institute (RENCI) and TeraGrid. We utilized the iRODS system [27] to store results directly from our workflows into secure storage. The biological software and databases used were: InterProScan (version 4.4), InterPro databases (version 23), UniProt (release 15.9), PSI-BLAST (v6.1) [28], TCDB (accessed October, 2009), NCBI ftp site for the *E. coli* and human datasets (accessed February, 2009).

Conclusions

The purpose of MotifNetwork is to rapidly identify functional characteristics of large numbers of protein sequences and put the results of this identification into a form amenable for further analysis. From the biological discovery point of view, it is seen that the domain based approach has special value in creating phylogenies, discovering novel proteins, and identifying potential drug targets in infectious pathogens. The combination of domain-based analysis methods coupled with MotifNetwork should permit researchers to more accurately and more efficiently process large sets of proteins to elucidate biologically significant analyses across species such as those illustrated in this paper with the three case studies.

The usefulness of these domain-based techniques, however is coupled to having results in a timely manner. This speed requirement is being addressed by MotifNetwork. The software engineering challenges for the deployed workflows and services are many. Generally identified scientific workflow issues include

- parameter-rich functions. How to handle data assembly and disassembly.
- access to varied and large amounts of computing; often through a grid execution model.
- workflow evolution. Once created, researchers needs require customized capabilities in their workflows

From the data management perspective, the main challenges are

- exponential growth of biological databases
- disparity of data representation formats
- data semantics and provenance, data quality and versioning control of datasets.

Authors' contributions

JLT designed and implemented the workflows and associated grid-services and developed the parallel applications. GR performed the biological analysis and developed analysis tools. EJ provided overall leadership and direction. All authors read and approved of this manuscript.

References

1. Orengo CA, Thornton JM: **Protein families and their evolution-a structural perspective.** *Annu Rev Biochem* 2005, **74:**867-890.
2. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes.** *J Mol Biol* 2001, **310:**311-325.

3. Fong JH, Geer LY, Panchenko AR, Bryant SH: **Modeling the evolution of protein domain architectures using maximum parsimony.** *J Mol Biol* 2007, **366:**307-315.

4. Björklund ÅK, Ekman D, Light S, Frey-Skott J, Elofsson A: **Domain rearrangements in protein evolution.** *J Mol Biol* 2005, **353:**911-923.

5. Song N, Sedgewick RD, Durand D: **Domain architecture comparison for multidomain homology identification.** *J Comp Biol* 2007, **14:**496-516.

6. Lin K, Zhu L, Zhang D-Y: **An initial strategy for comparing proteins at the domain architecture level.** *BMC Bioinformatics* 2006, **22:**2081-2086.

7. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A: **Arrangements in the modular evolution of proteins.** *Trends in Biochemical Sciences* 2008, **33:**444-451.

8. Tasneem A, Iyer L, Jakobsson E, Aravind L: **Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop in channels.** *Genome Biol* 2005, **6:**R4.

9. Bocquet N, Prado de Carvalho L, Cartaud J, Neyton J, Le Paupon C, Taly A, Grutter T, Chamgeux J-P, Corringer P-J: **A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family.** *Nature* 2007, **445:**116-119.

10. Hilf RJC, Dutzler R: **X-ray structure of a prokaryotic pentameric ligand-gated ion channel.** *Nature* 2008**:**375-379.

11. Tilson JL, Blatecky A, Rendon G, Ger M-F, Jakobsson E: **Genome-Wide Domain Analysis using Grid-enabled Workflows.** In *7th IEEE International*

Conference on Bioinformatics and Bioengineering (BIBE 2007); Harvard Medical School, Boston, MA. IEEE; 2007: 872-879.

12. Tilson JL, Rendon G, Ger M-F, Jakobsson E: **MotifNetwork: A Grid-enabled Workflow for High-throughput Domain Analysis of Biological Sequences: Implications for annotation and study of phylogeny, protein interactions, and intraspecies variation.** In *7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE 2007); Oct 14-17; Harvard Medical School, Boston, MA*. IEEE; 2007: 620-627.
13. Zdobnov EM, Apweiller R: **InterProScan - an integration platform for the signature-recognition methods in interPro.** *Bioinformatics* 2001, **17:**847-848.
14. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37:**D211-215.
15. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology; Menlo Park, CA*. AAAI Press; 1994: 28-36.
16. Bailey TL, Michael G: **Combining evidence using p-values: application to sequence homology searches.** *Bioinformatics* 1998, **14:**48-54.
17. Saier MH, Jr., Tran CV, Barabote RD: **TCDB: the Transporter Classification Database for membrane transport protein analyses and information.** *Nucl Acids Res* 2006, **34:**D181-186.
18. Rendon G, Ger M-F, Kantorovitz R, Natarjan S, Tilson J, Jakobsson E: **Understanding the "Horizontal Dimension" of Molecular Evolution to**

Annotate, Classify, and Discover Proteins with Functional Domains. *Journal of Computer Science and Technology* 2010, **25:**82-94.

19. Shames SR, Auweter SD, Finlay BB: **Co-evolution and exploitation of host cell signaling pathways by bacterial pathogens.** *Int J Biochem Cell Biol* 2009, **41:**380-389.

20. Gropp W, Lusk E, Skjellum A: *Using. MPI: Portable parallel programming with the message passing interface.* MIT Press; 1994.

21. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Idekerm T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.

22. Tilson JL, Kandaswamy G, Rendon G, Jakobsson E: **MotifNetwork: High Throughput Determination of Evolutionary Domain Networks.** In *International Conference on Bioinformatics & Computational Biology (BIOCOMP 2009); Las Vegas, NV*. Edited by Arabnia HR, Yang MQ. CSREA Press; 2009: 221-227.

23. Oinn T, Greenwood M, Addis M, Alpdemir N, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin D, et al: **Taverna: lessons in creating a workflow environment for the life sciences.** *Concurrency and Computation: Practice and Experience* 2006, **18:**1067-1100.

24. Foster I, Kesselman C: In *The Grid 2: Blueprint for a New Computing Infrastructure.* CA: Morgan-Kaufman; 1999

25. Foster I, Kesselman C, Tuecke S: **The anatomy of the grid: Enabling scalable virtual organizations.** *Int J Supercomputing App* 2001, **15**.

26. Foster I: **Globus toolkit version 4: software for service oriented systems.** In *IFIP International Conference on Network and Parallel Computing*. Springer-Verlag; 2006: 2-13.

27. Rajasekar A, Wan M, Moore R, Schroeder W: **A prototype rule-based distributed data management system.** In *High Performance Distributed Workshop (HPDC) on Next Generation Distributed Data Management; Paris, France*. 2006

28. Altshul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSIBLAST: a new generation of protein search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

Acknowledgements

We gratefully acknowledge support from the National Science Foundation under Grant Nos. 0835651 and 0835718. We also wish to thank the Renaissance Computing Institute for providing access to large-scale computing platforms and Bradley Viviano (RENCI) for essential cluster, security, and Globus GT4 support.

"Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation."

Figures

Figure 1 - Case Study 1. Comparison of four methods of domain-based homology

TCDB_id	LIC Subfamily	Domain Composition Pattern	dp_true_positives	dp_tot	dp_weight	idf_true_positives	idf_tot	idf_weight	cacf_true_positives	cacf_tot	cacf_weight	jaccard_true_positives	jaccard_tot	jaccard_weight
1.A.9.1	ACH, ACR	IPR002394 IPR006029 IPR006201 IPR006202 IPR018000	795	795	[1.00-0.7069]	764	764	[1.0-0.8890]	708	708	[1.0-0.9090]	708	708	[1.0-0.8333]
1.A.9.2	5HT3	IPR006029 IPR006201 IPR006202 IPR008132 {IPR008133 | IPR008134} IPR018000	37	37	[1.0-0.3031]	22	22	[1.0-0.8333]	25	25	[1.0-0.8000]	17	17	[1.0-1.0]
1.A.9.3.1	GLRA	IPR006028 IPR006029 IPR006201 IPR006202 IPR008127 { IPR008128 | IPR008129 |IPR008130 } IPR018000	47	47	[1.0-0.2033]	47	47	[1.0-0.7666]	46	46	[1.0-0.8571]	46	46	[1.0-0.7500]
1.A.9.3.1	GLRB	IPR006029 IPR006201 IPR006202 IPR008060 IPR018000	19	19	[1.0-0.8785]	14	14	[1.0-0.9012]	13	13	[1.0-1.0]	13	13	[1.0-1.0]
1.A.9.3.2	HCLA	IPR006028 IPR006029 IPR006201 IPR006202	7	27	[1.0-1.0]	7	27	[1.0-1.0]	7	27	[1.0-1.0]	7	27	[1.0-1.0]
1.A.9.3.3	HCLB	IPR006028 IPR006029 IPR006201 IPR006202 IPR018000	30	259	[1.0-1.0]	30	259	[1.0-1.0]	30	259	[1.0-1.0]	30	259	[1.0-1.0]
1.A.9.4	GLUCL	{ IPR006028 IPR006029 } IPR006201 IPR006202 IPR015680 IPR018000	152	152	[1.0-0.814]	89	88	[1.0-0.9133]	87	87	[1.0-0.9090]	86	86	[1.0-0.8571]
1.A.9.5.1 1.A.9.5.2	GBRB	IPR002289 IPR006028 IPR006029 IPR006201 IPR006202 IPR018000	143	143	[1.0-0.8973]	130	130	[1.0-0.9013]	130	130	[1.0-0.9090]	130	130	[1.0-0.9130]
1.A.9.5.2	GBRAP	IPR004241	470	470	[1.0-0.3333]	470	470	[1.0-0.5215]	470	470	[1.0-0.6666]	470	470	[1.0-0.2500]
1.A.9.5.2	GBRA2	IPR001390 IPR005432 IPR006028 IPR006029 IPR006201 IPR006202 IPR018000	77	77	[1.0-0.2033]	113	113	[1.0-0.7949]	113	113	[1.0-0.8333]	113	113	[1.0-0.7143]
1.A.9.6	MOD-1, LGC-53, LGC-40,	IPR006028 IPR006029 IPR006201 IPR006202 {IPR018000}	5	286	[1.0-1.0]	5	286	[1.0-1.0]	5	286	[1.0-1.0]	5	286	[1.0-1.0]
1.A.9.7	EXP_1	IPR006028 IPR006029 IPR006201 IPR006202 IPR018000	3	259	[1.0-1.0]	3	259	[1.0-1.0]	3	259	[1.0-1.0]	3	259	[1.0-1.0]
1.A.9.8	GLVI	IPR006028 IPR006201 IPR006202	2	2	[1.0-1.0]	2	2	[1.0-1.0]	2	2	[1.0-1.0]	2	2	[1.0-1.0]
1.A.9.9	ELIC	IPR006201 IPR006202	25	150	[1.0-1.0]	25	150	[1.0-1.0]	25	150	[1.0-1.0]	25	150	[1.0-1.0]
1.A.9	other LIC related			381			473			529			538	

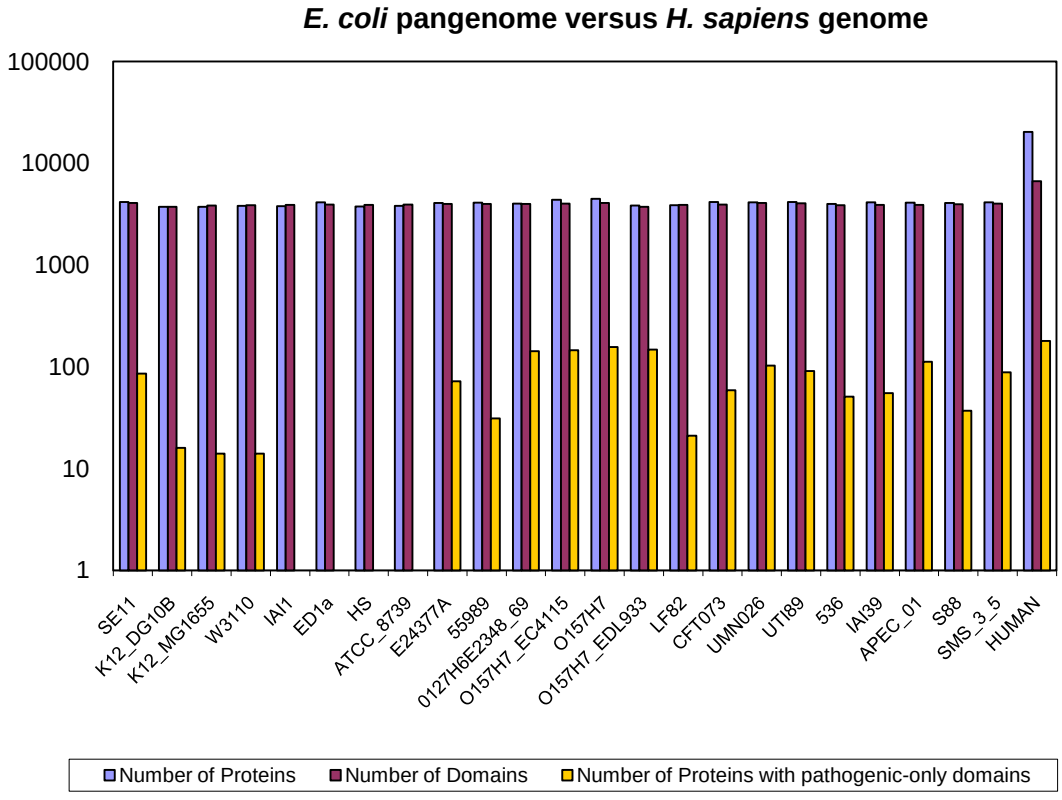
Results of four approaches to domain-based functional homology of the LIC family (1.A.9 TCDB). White-colored columns reports results for the distinct-partners function. Purple columns report the inverse document coefficient function. Green columns report the cumulative association coefficient function. Yellow columns report results of using the Jaccard function. For composition patterns, optional domains are designated using [1] and while | denote the selection of only one option.

Figure 2 - Case Study 2. Prokaryotic homologs of the LIC family

2VL0	NP_927143.1	YP_002466490.1	YP_781543.1	EDM81546.1	YP_667426.1
CAM74622.1	NP_948199.1	YP_002484938.1	YP_898128.1	YP_001891319.1	YP_001403577.1
EAM48767.1	YP_001121761.1	YP_002599689.1	EDY03702.1	YP_532370.1	YP_002168189.1
EAR18354.1	YP_001226021.1	YP_002655067.1	YP_002374163.1	EDN69508.1	YP_678409.1
EAU72723.1	YP_001404166.1	YP_170294.1	YP_763066.1	EDN69487.1	YP_513219.1
EAW36637.1	YP_001427905.1	YP_419787.1	EDX19711.1	YP_001992176.1	YP_001805637.1
EAZ91136.1	YP_001677092.1	YP_486367.1	YP_002170357.1	YP_569928.1	EDM72298.1

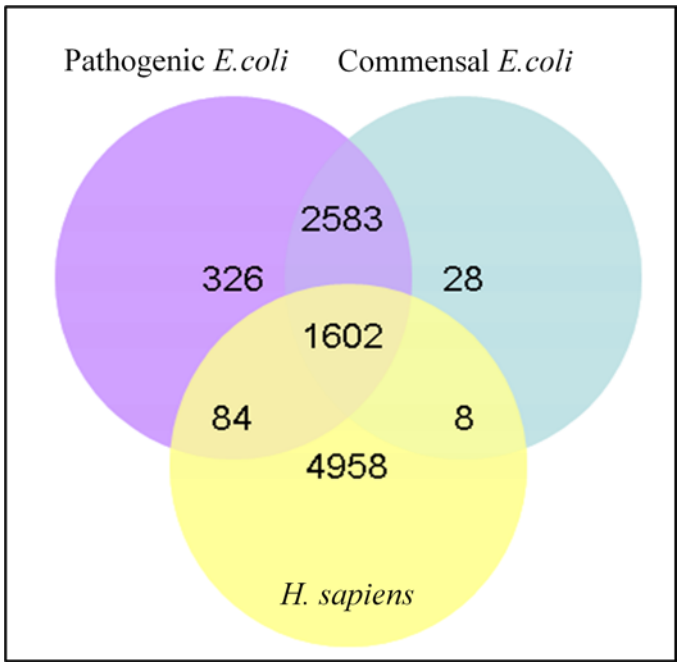
A list of likely prokaryotic homologs of the LIC family (1.A.9 TCDB) identified by GenBank id number. Highlighted in blue are two confirmed prokaryotic members of the LIC family: GLVI (GenBank: NP_927143.1), and ELIC (PDB: 2VL0).

Figure 3 - Case Study 3. Comparison of the total protein/domain content for 23 genomes of *E. coli* and *H. Sapiens*



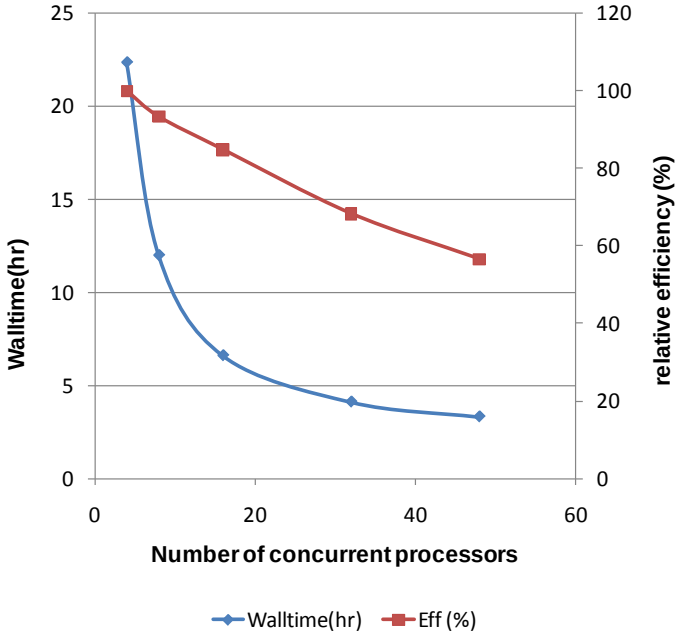
Histogram, on a logarithmic scale, that displays the total counts of proteins, Interpro domains, and proteins with only pathogenic domains for each strain of *E. coli* and for *H. sapiens*.

Figure 4 - Case Study 3. Comparison of the total protein/domain content for 23 genomes of *E. coli* and *H. sapiens.* Venn diagram



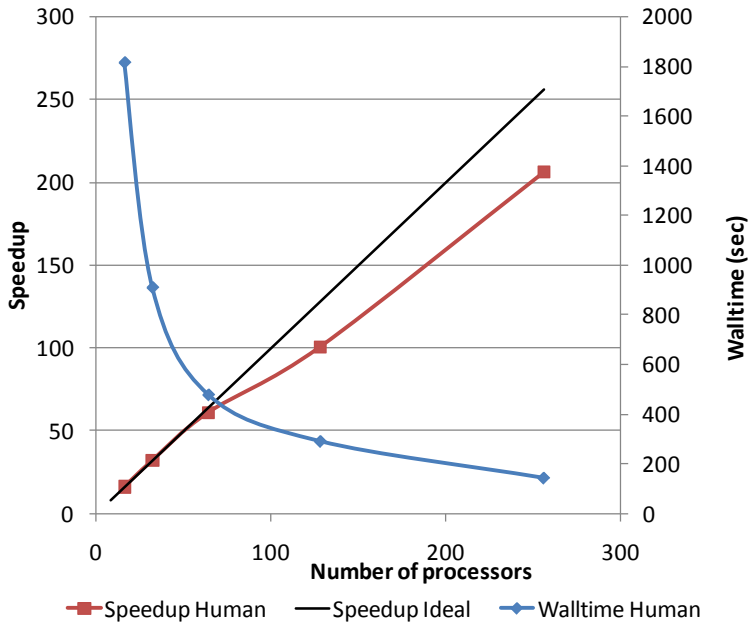
Venn diagram describing the domain distribution of pathogenic and commensal *E. coli* and *H sapiens*. 1602 domains are contained in all three classes of organisms. Commensal *E. coli* contains only 28 domains distinct from its pathogenic counterparts whereas 326 domains in pathogenic *E. coli* are unique. This group of pathogen-unique domains comprises a pool of potential targets for inhibiting pathogenicity in these microbes.

Figure 5 - Workflow performance as a function of ensemble width



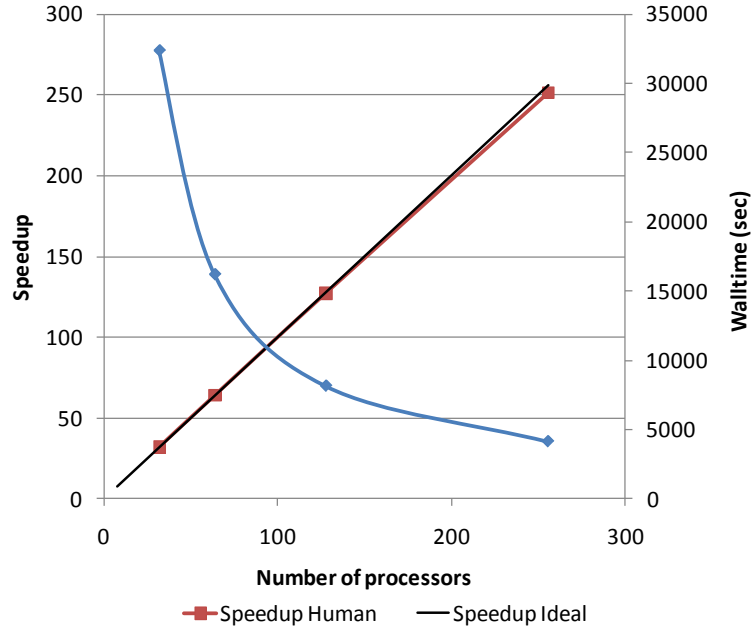
Measure walltime (hr) versus the maximum number of available concurrent processors. The workflow can use fewer than the maximum depending on the total number of jobs that comprise the ensemble. The data are for processing of *E. coli* (W3110)

Figure 6 - Parallel performance for the ScoreMatrix program



Measured speedup (SU) and walltime (sec) for the ScoreMatrix step of a typical MotifNetwork analysis. Quantities are versus the number of applied processors. Data are for *H. sapiens*.

Figure 7 - Parallel performance for the WebMatrix program



Measured speedup (SU) and walltime (sec) for the WebMatrix step of a typical MotifNetwork analysis. Quantities are versus the number of applied processors. Data are for *H. sapiens.*

Tables

Table 1 - Interpro domains of the LIC family

IPR001390 Gamma-aminobutyric-acid A receptor, alpha subunit
IPR002289 Gamma-aminobutyric-acid A receptor, beta subunit
IPR002394 Nicotinic acetylcholine receptor, N-terminal
IPR004241 MAP1_LC3
IPR005432 Gamma-aminobutyric-acid A receptor, alpha 2 subunit
IPR006028 Gamma-aminobutyric acid A receptor
IPR006029 Neurotransmitter-gated ion-channel transmembrane region
IPR006201 Neurotransmitter-gated ion-channel
IPR006202 Neurotransmitter-gated ion-channel ligand-binding
IPR008060 Glycine receptor beta
IPR008127 Glycine receptor alpha
IPR008128 Glycine receptor alpha1
IPR008129 Glycine receptor alpha2
IPR008130 Glycine receptor alpha3
IPR008132 5-hydroxytryptamine 3 receptor
IPR008133 5-hydroxytryptamine 3 receptor, A subunit
IPR008134 5-hydroxytryptamine 3 receptor, B subunit
IPR015680 Glutamate-Gated Chloride Channel
IPR018000 Neurotransmitter-gated ion-channel, conserved site