BMC Bioinformatics

Research article

# Species-specific protein sequence and fold optimizations

Michel Dumontier[1,2], Katerina Michalickova[1,2] and
Christopher WV Hogue*[1,2]

Address: [1]Department of Biochemistry, University of Toronto, Toronto, Ontario, M5S 1A8, Canada and [2]Samuel Lunenfeld Research Institute,
Mount Sinai Hospital, 600 University Ave., Toronto, Ontario, M5G 1X5 Canada

E-mail: Michel Dumontier - micheld@mshri.on.ca; Katerina Michalickova - katerina@mshri.on.ca;
Christopher WV Hogue* - hogue@mshri.on.ca

*Corresponding author

Published: 17 December 2002

*BMC Bioinformatics* 2002, **3**:39

Received: 11 July 2002
Accepted: 17 December 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/39

## Abstract

**Background:** An organism's ability to adapt to its particular environmental niche is of fundamental importance to its survival and proliferation. In the largest study of its kind, we sought to identify and exploit the amino-acid signatures that make species-specific protein adaptation possible across 100 complete genomes.

**Results:** Environmental niche was determined to be a significant factor in variability from correspondence analysis using the amino acid composition of over 360,000 predicted open reading frames (ORFs) from 17 archae, 76 bacteria and 7 eukaryote complete genomes. Additionally, we found clusters of phylogenetically unrelated archae and bacteria that share similar environments by amino acid composition clustering. Composition analyses of conservative, domain-based homology modeling suggested an enrichment of small hydrophobic residues Ala, Gly, Val and charged residues Asp, Glu, His and Arg across all genomes. However, larger aromatic residues Phe, Trp and Tyr are reduced in folds, and these results were not affected by low complexity biases. We derived two simple log-odds scoring functions from ORFs ($C_G$) and folds ($C_F$) for each of the complete genomes. $C_F$ achieved an average cross-validation success rate of 85 ± 8% whereas the $C_G$ detected 73 ± 9% species-specific sequences when competing against all other non-redundant $C_G$. Continuously updated results are available at [http://genome.mshri.on.ca].

**Conclusion:** Our analysis of amino acid compositions from the complete genomes provides stronger evidence for species-specific and environmental residue preferences in genomic sequences as well as in folds. Scoring functions derived from this work will be useful in future protein engineering experiments and possibly in identifying horizontal transfer events.

## Background

An organism may increase its fitness in some range of environmental conditions through evolution. Fundamental to the survival of cells is the ability to modulate fluctuations in external osmotic and atmospheric pressure, temperature and pH via the acquisition or development of advantageous molecular mechanisms [1–4]. These mechanisms include the uptake of small molecules, osmolytes or metals via transporters as found for increased iron uptake allowing enhanced growth of *Pasteurella multocida* [5] and in the accumulation of high concentrations of the stabilizing K+ among halophiles [6]. Other mechanisms include modification of the atomic [7] and residue [8] composition of proteins, or the acquisition of environmental adaptive genes via lateral gene transfer as was likely the case for the thermophilic bacteria *Thermotoga maritima* [9] and archaea *Solfolobus solfataricus P2* [10]. In other cases, the gene duplication events augment the ability of an organism to adapt to extreme environments by expanding specific protein families including additional stress response and damage control genes that provide increased protection for the radiation resistant bacteria *Deinococcus radiodurans* [11,12]. Interestingly, in symbionts such as *Buchnera sp. APS* [13], *Agrobacterium tumefaciens* [14] and *Sinorhizobium meliloti* [15], shared genetic material may increases overall fitness, but this effectively results in the loss of redundant genes and imposes host-symbiont dependencies. In other organisms completely new and innovative mechanisms are required for adapting to the most extreme of environments.

In adaptation to the most extreme environments, it is expected that the protein complement also possesses the organism's adaptive property [6]. For instance, hyperthermophilic proteins must not only be functional, but optimized towards the host's extremely hot (>80°C) physical environment. Although *in vivo* protection factors have been identified that can stabilize proteins *in vitro* at high temperatures [1] and chaperone proteins can help refold misfolded proteins and prevent aggregation [16–18], the majority of foreign proteins cloned and expressed in *E. coli* retain all of the native enzyme's biochemical properties, including proper folding, thermostability and optimal activity consistent with the organism's optimal growth temperatures [19–21]. Thus, it is likely that sequence optimizations are required to ensure protein activity and folding in organisms whose growth conditions might otherwise adversely affect proteins.

Researchers have studied complete or partial genomes using bioinformatics in addition to the traditional comparative sequence-structure and structure-function mutation studies to identify stability factors. Recent studies of complete or partial genomes have identified sequence-based correlations between organisms using amino acid compositions. Lobry demonstrated the correlation between G+C content and codon usage across bacterial sequences [22] and G+C content and amino acid composition correlations have been extended to 25 complete genomes [8]. Moreover, codon usage and amino acid preferences for thermophiles are well established and have been extended to complete genomes [23–26]. However, these generalizations do not necessarily agree with comparative sequence-structure studies. Comparative studies often exploit sequence or structure based alignments to determine similarities and differences. Investigation of thermostability factors across 10 organisms including psychrophiles (cold-tolerant), mesophiles to hyperthermophiles with triosephosphate isomerase failed to identify significant correlations of composition with thermostability [27]. Further uncertainty arises from indications that different protein families adapt to temperature conditions by different sets of structural mechanisms [28]. How then to unify amino acid composition preferences with species-specific structural adaptations?

Algorithms have been designed to predict certain protein features primarily from sequence composition including low complexity regions [29], transmembrane segments [30], signal peptides [31], coiled-coils [32], secondary structure elements [33], structural classes [34], hydrophobicity [35], sub-cellular location [36] and have been used to increase remote sequence similarity searching [37,38]. Moreover, genomic base content has been used to predict open-reading frames and in-site splicing [39–41]. However, no algorithms have been designed to explore adaptation of proteins to their host environment, especially in a species-specific manner.

Species-specific adaptive optimizations might be expected to be subtle and hard to find in any individual sequence, yet sufficiently common across the bulk of genomic proteins that they may be detected using statistical methods. We demonstrate here that such subtle adaptive optimizations do exist in many individual organisms and that these can be extracted. We derive species-specific protein sequence and fold scoring functions from residue preferences found in predicted open reading frames and conservative structural models. The resulting scoring functions are effective in amino acid composition species-specific protein sequence and fold detection.

## Results and Discussion

### *Principal Components Analysis*

Principal Components Analysis (PCA) was performed with the amino acid compositions of the entire set of protein coding regions from each of the complete genomes (Figure 1). PCA transforms a number of (possibly) correlated variables into uncorrelated variables called principal components that account for the variance in the dataset (see [http://www.statsoftinc.com/textbook/stfacan.html] for brief overview). The analysis involves plotting the original variables to the principal components (factor loadings) and can be interpreted as correlation coefficients (Figure 1B,1D). Factor loadings of = 0.6 are considered to be strong correlations. Simultaneously, a correspondence of the mean genome amino acid compositions to the principal components may be observed in order to observe genomic usage or preference that appear to correlated factor loadings (Figure 1A,1C).

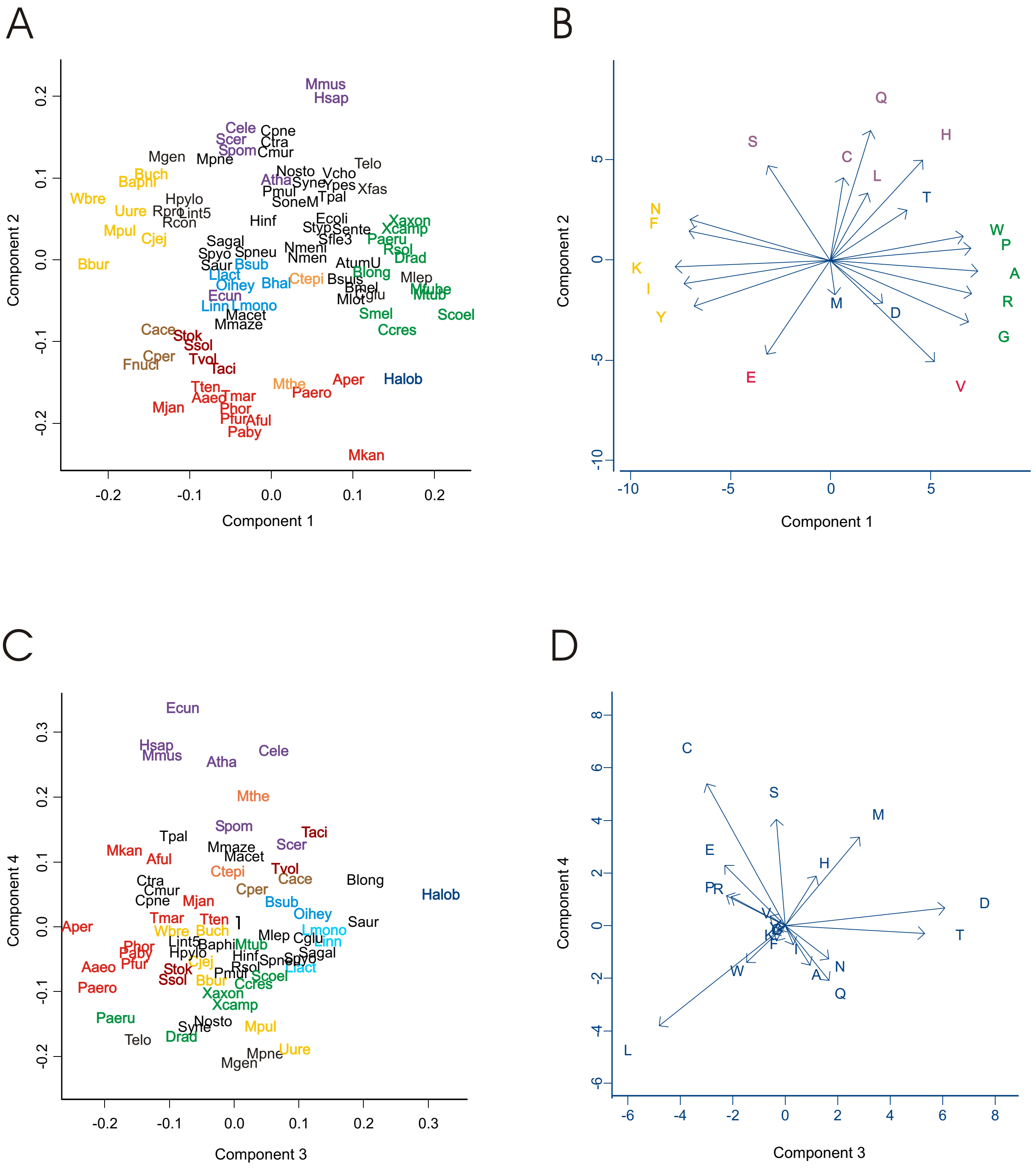


**Figure 1**

**Principal Components Analysis** Plots of principal components 1, 2 (A, B) and 3, 4 (C, D) obtained from the amino acid composition of all their predicted open-reading frames as they correspond to the mean composition of the complete genomes (A, C) and their amino acid factor loadings (B, D). GC poor genomes (yellow), GC rich genomes (green), hyperthermophiles (red), thermophiles (orange), thermo-acidophiles (red-brown), solventogens (brown), alkalophiles (blue), extreme halophile (navy), and eukaryotes (purple). Note that there is only one genome representative for any cluster of strains or variants (i.e. Ecoli, EcoliE and EcoliH are all represented by Ecoli). In C, all remaining organisms are clustered around the number 1.

The most significant principal component accounted for 47.5% of the variance and showed a strong correlation to DNA base pair content (94%). The left of this component corresponds to low GC organisms such as *buchnera sp.* (~27%), Mpul (~27%), Bbur (29%), Uure (26%), Wbre (23%) whereas the right of the component corresponds to high GC organisms including Mtub (66%), various plant pathogens (*Xanthomonas sp.*, Mloti), soil bacterium Scoel (72%) and radiation-resistant Drad (66.6%) (Figure 1A). Strong correlations also exist between the first component with several of the factor loadings (Figure 1B). The correlated factor loadings have either [G|C] or [A|T] in the first two codon positions for some codon. The effect for the standard codon table is that GC rich codons [C|G] [C|G] [X] encode amino acids Pro, Arg, Gly, Ala, Trp and GC poor codons [A|T] [A|T] [X] encoding Phe, Leu, Ile, Asn, Lys, and Tyr (as well as Met and 2 stop codons). This is in agreement with a previous report [8]. Consequently, genomic GC content will to a large extent determine amino acid usage as well as the choosing between small hydrophobic residues Ala/Gly or Ile, positively charged residues Arg or Lys, and large hydrophobic residues Trp or Tyr/Phe.

The second largest principal component accounts for 15.5% of the variance and appears to correspond to the environmental niche (Figure 1A). Hyperthermophiles (Mkan, Paby, Pfur, Phor, Aful, Aaeo, Tmar, Tten, and Mjan), thermophile Mthe, extreme halophile Halo, thermo-acidophiles (Taci, Tvo, Ssol, Stok), and solventogenic bacteria (Cace, Cper and Fnucl) correspond strongly to weakly, respectively, to component 2. Strong correlations to this component exist for Glu and Val, although opposite correlations exist for Gln, His, Thr, Ser and Cys, thereby suggesting the preferential usage of these amino acids by those organisms. A discussion regarding amino acid preferences for hyperthermophiles can be found elsewhere [8]. Eukaryotes (Hsap, Mmus, Scer, Cele and Atha), with the exception of the obligate intracellular eukaryote parasite Ecun, have a strong, but opposite correspondence to component 2. These cluster with chlamydias/chlamydophilas (Cmur, Ctra, Cpne) and the inverse correspondence also indicates a significant increase in the genomic amino acid usage of Gln, His, Thr, Ser, and Cys, and decrease of Glu or Val. Interestingly, plant pathogens (Xaxon, Xcamp, Mloti, Rsol, Xfas, AtumC and AtumU), moderate halophiles and alkalophiles (Bsub, Bhal, Linn, Lact, Lmono, Oihey), and most human pathogens do not correspond to this component and have an average composition with regards to these amino acids. The distribution of organisms across this component does not appear to correspond to discrete groupings of organisms that share similar environmental niches, but rather to a 'continuum of lifestyles' [26]. However, unlike previous studies that report correlations of this second principal component with growth temperatures [8,26], our results seem to indicate that this component is likely to correlate to a more complex phenomenon that incorporates growth temperature as well as other physical factors, possibly pH and solvent.

Components 3 and 4 are also significant factors in this multivariate analysis and these account for 10.3% and 7.4% of the variance. We have not determined a measurable factor that can be directly correlated to these components, but they also appear to correspond to environmental niche. However, we see species-specific preferences for Leu, Cys, Asp, Thr, Ser, and to a lesser extent Glu, Gln, His and Met residues (Figure 1D). Component 3 strongly corresponds to several hyperthermophiles, but inversely corresponds to the extreme halophile *Halobacterium*, human pathogen Saur, gastro-intestinal tract colonizer Blong, and moderate halophiles and alkalophiles. *Halobacterium*'s increased Asp usage is clearly consistent with its adaptation to intracellular and environmental conditions [42], although it differs to the hyperthermophile preference for the larger, negatively charged Glu. Component 4 has strong correspondence to the eukaryotes (Ecun, Hsap, Mmus, Atha, Cele) that correlates to Cys and Ser.

Taken together, the results from the principal component analysis suggest that amino acids that vary significantly among and between species are due to a large extent to environmental conditions.

### *Amino acid composition dendrogram*

To compare organism amino acid composition, we performed hierarchical clustering using the complete linkage method with distances computed using the Euclidean metric on a dataset that consisted of the mean percent amino acid composition from all predicted open-reading frames for each of the 100 organisms (Figure 2). This method generates clusters of organisms with a similar mean composition across all 20 amino acids that are maximally separated by using the farthest neighbours. The resulting dendrogram presents three large branches within 10 Euclidean difference units. The upper branch clusters genomes with low GC content (yellow), the mid branch clusters mid GC genomes and the lower branch clusters high GC content genomes (green). A feature of clustering by amino acid composition is that phylogenetically related organisms are not necessarily proximate neighbours. For instance, Hsap and Mmus are clustered together, but are separated by a significant distance from Spom, Scer, Atha and Cele as well as the eukaryote Ecun. Oddly, Ecun clusters closely to hyperthermophilic archae Aful and thermophilic Mthe and more distantly to a cluster comprised of hyperthermophilic bacteria Aaeo and Tmar and archae Paby, Phor and Pfur. However, this organism is not reported to have thermophilic qualities [43]. In another

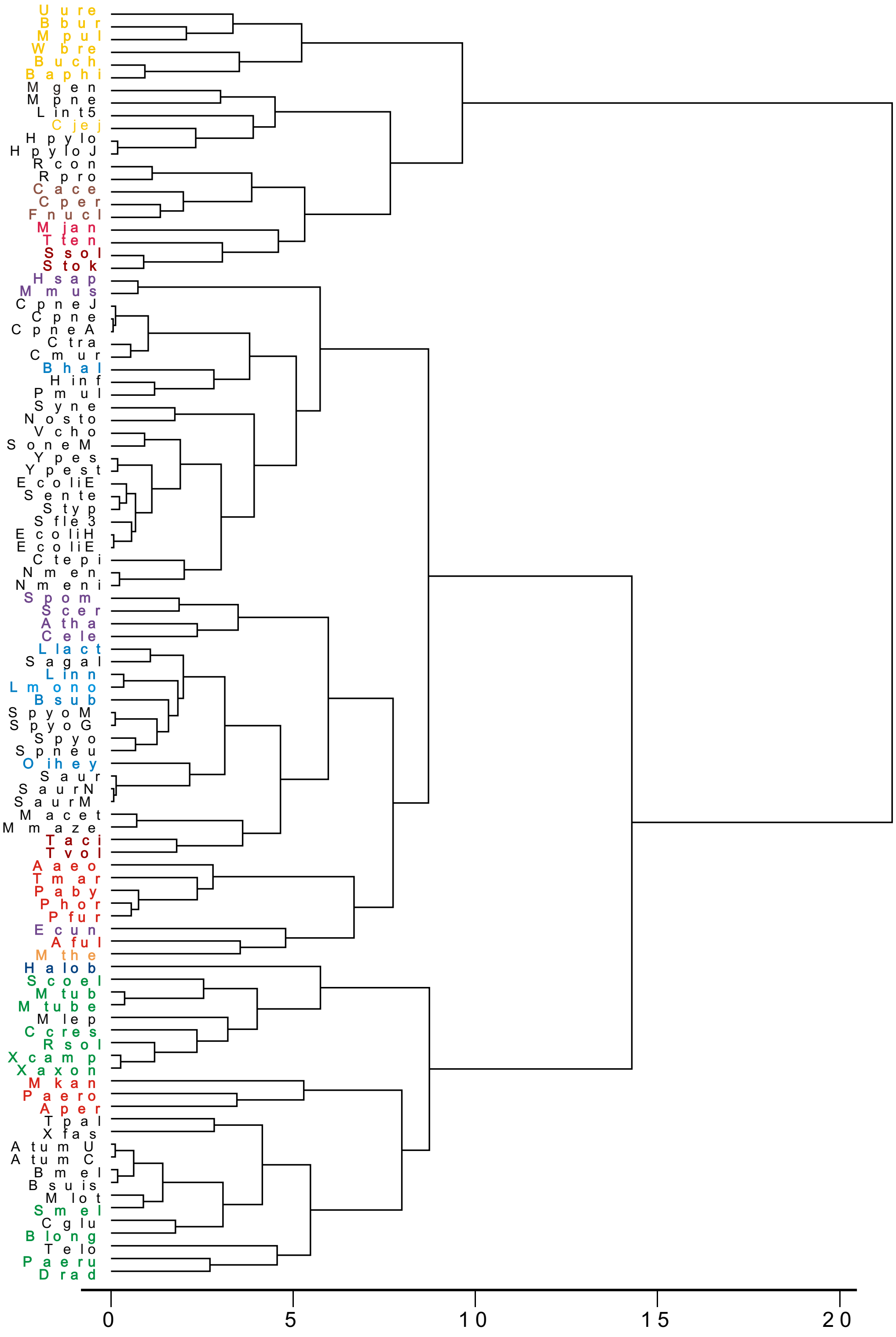


**Figure 2**

**Amino acid composition dendrogram** Amino acid composition dendrogram obtained from clustering the average amino acid composition of each genome. Hyperthermophiles (red), thermophiles (orange), (thermo)-acidophiles (brown), solventogens (brown), alkalophiles (blue), extreme halophile (blue) and eukaryotes (purple). The scale represents Euclidian distance.

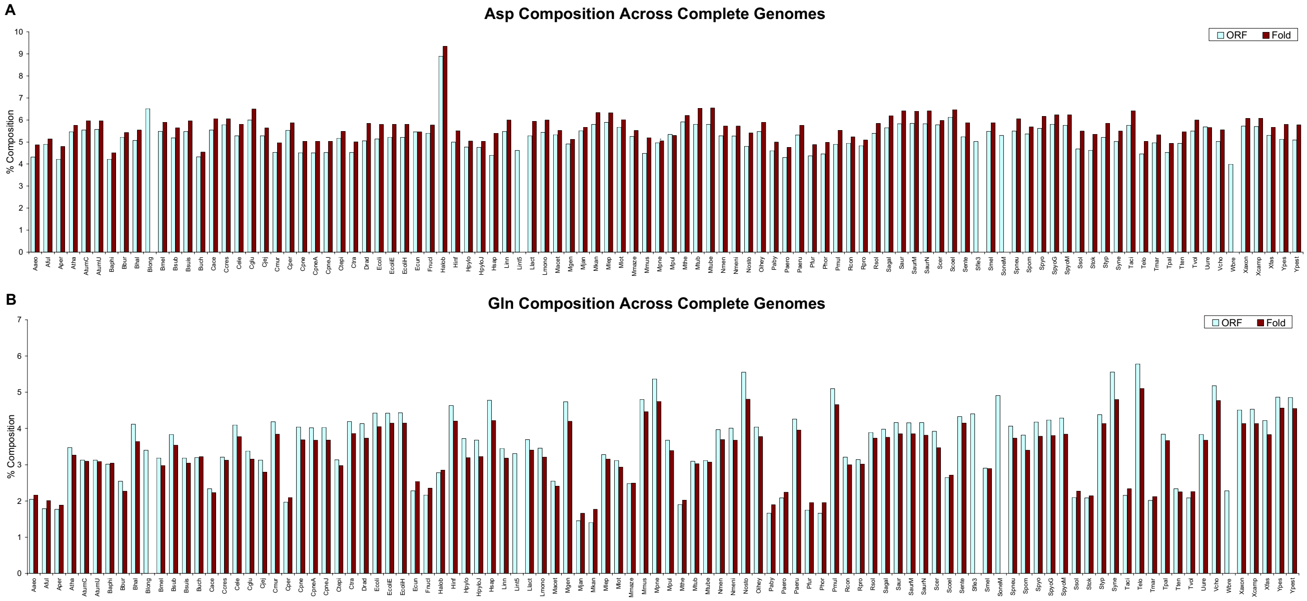


**Figure 3**

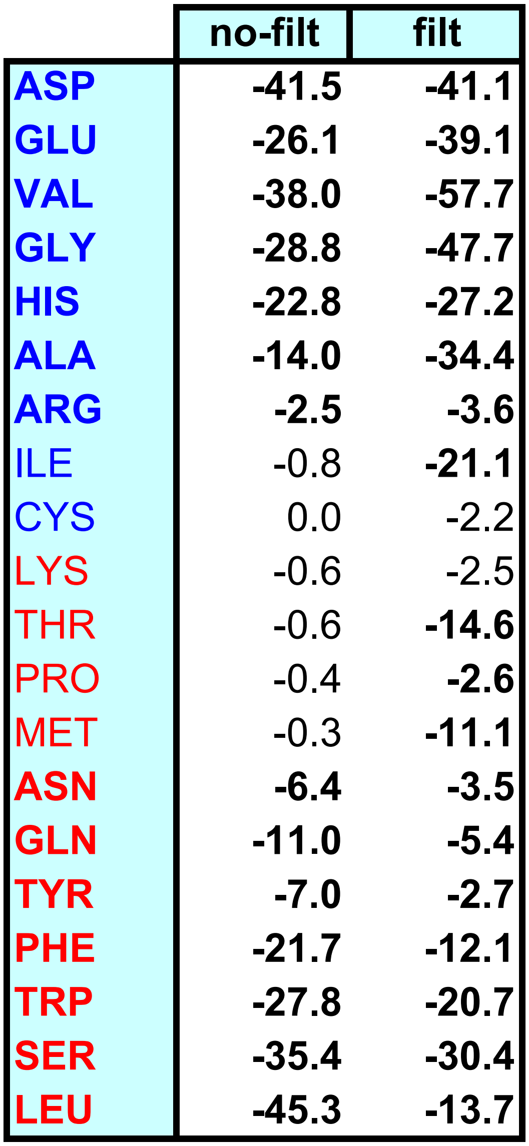
**Comparison of ORF and Fold composition from complete genomes** Amino acid composition from predicted open-reading frames (ORF, blue) and fold regions (Fold, red) of Asp (A) and Gln (B) for each complete genome. Bold values indicate significantly large preferences for (positive) or against (negative) certain residues.

case, hyperthermophiles Aper, Paero and Mkan are clustered together, indicating that organisms that are less phylogenetically related may form tight clusters of organisms that live in similar environments. These results significantly extend previous composition-based dendrograms [8], but differ significantly from other attempts to generate genome-based dendrograms [44,45].

### *Fold residue preferences*

In order to address the question of whether the amino acid composition of ORFs were different that of folds as well as whether fold composition was species-specific, we generated over 57,000 conservative domain-based structure models for 95 genomes (see materials and methods). Amino acid compositions were computed across all protein coding regions for each complete genome using either genomic sequences for a given organism ($C_G$) or fold ($C_F$) for the purpose of identifying species-specific as well as pan-specific fold composition bias. Furthermore, excluded indels residues from the modeling exercise comprised <2% of all residues and these exhibited normal insertion or loop compositions richer in Pro and Gly, but poorer in the small hydrophobic residues. Figure 3 illustrates one case in which the mean composition of Asp is unvarying across all genomes, with the single exception of the extreme halophile Halobacterium. Moreover, we observe a significant increase in Asp residues in the fold regions as compared to the predicted ORF (t-test: $p < 10^{-38}$). Figure 3 also illustrates a case in which the mean composition of Gln varies significantly across the genomes. Virtually all genomes show a decrease of Gln ($p < 10^{-11}$) in the fold regions, with the startling exception of all thermophiles as well as Cper, Ecun, Halob, Scoel, Buchn, Fnucl and Mmaze. Although the mean composition of Gln is significantly lower ($p < 10^{-24}$) in these thermophiles than the other genomes, the increase of Gln in the fold is a surprising finding given that amidated residues are susceptible to deamidation at high temperatures [46,47]. However, others have reported that polar residues such as Gln are significantly reduced on the surface of thermophilic intracellular proteins as compared to their mesophilic counterparts, likely reducing the possibility of damaging deamidation reactions [48].

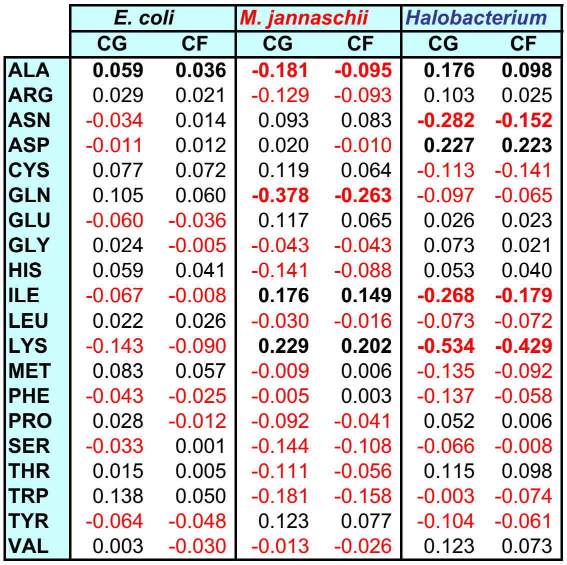
We found that small hydrophobic residues Ala, Gly and Val as well as charged residues Asp, Glu, His and Arg are consistently increased in the fold regions across all organisms (Figure 4). Furthermore, we observed a significant decrease of amidated residues Asn and Gln as well as larger aromatic residues Phe, Trp, and Tyr, as well as Leu and Ser in the fold regions. It is possible that smaller residues in fold regions allow better packing of the core whereas charged residues are utilized for stabilizing electrostatic interactions including salt bridges. In order to exclude the possibility that our results may be biased due to low compositional complexity of ORF or fold regions, we applied



| | no-filt | filt |
|---|---|---|
| **ASP** | **-41.5** | **-41.1** |
| **GLU** | **-26.1** | **-39.1** |
| **VAL** | **-38.0** | **-57.7** |
| **GLY** | **-28.8** | **-47.7** |
| **HIS** | **-22.8** | **-27.2** |
| **ALA** | **-14.0** | **-34.4** |
| **ARG** | **-2.5** | **-3.6** |
| ILE | -0.8 | **-21.1** |
| CYS | 0.0 | -2.2 |
| LYS | -0.6 | -2.5 |
| THR | -0.6 | **-14.6** |
| PRO | -0.4 | **-2.6** |
| MET | -0.3 | **-11.1** |
| **ASN** | **-6.4** | **-3.5** |
| **GLN** | **-11.0** | **-5.4** |
| **TYR** | **-7.0** | **-2.7** |
| **PHE** | **-21.7** | **-12.1** |
| **TRP** | **-27.8** | **-20.7** |
| **SER** | **-35.4** | **-30.4** |
| **LEU** | **-45.3** | **-13.7** |

**Figure 4**
**ORF and Fold compositions are significantly different** Log of two-tailed paired t-test probabilities between ORF and fold amino acid mean compositions across all genomes, without filtering (no-filt) and with four filtering methods: transmembrane, coiled-coil, low-complexity and compositional bias (filt). Values of less than -2.5 indicate a significant difference.

transmembrane, coiled-coil, compositional bias and low complexity region filtering using the pfilt application from David T. Jones (1997) and found few deviations from these trends (Figure 4). Since a large number of our templates are obtained via crystallography experiments, we cannot rule out the possibility that the fold composition bias may reflect a composition that is more amenable to crystallographic structure determination.



| | *E. coli* | | *M. jannaschii* | | *Halobacterium* | |
|---|---|---|---|---|---|---|
| | CG | CF | CG | CF | CG | CF |
| ALA | **0.059** | **0.036** | **-0.181** | **-0.095** | **0.176** | **0.098** |
| ARG | 0.029 | 0.021 | -0.129 | -0.093 | 0.103 | 0.025 |
| ASN | -0.034 | 0.014 | 0.093 | 0.083 | **-0.282** | **-0.152** |
| ASP | -0.011 | 0.012 | 0.020 | -0.010 | **0.227** | **0.223** |
| CYS | 0.077 | 0.072 | 0.119 | 0.064 | -0.113 | -0.141 |
| GLN | 0.105 | 0.060 | **-0.378** | **-0.263** | -0.097 | -0.065 |
| GLU | -0.060 | -0.036 | 0.117 | 0.065 | 0.026 | 0.023 |
| GLY | 0.024 | -0.005 | -0.043 | -0.043 | 0.073 | 0.021 |
| HIS | 0.059 | 0.041 | -0.141 | -0.088 | 0.053 | 0.040 |
| ILE | -0.067 | -0.008 | **0.176** | **0.149** | **-0.268** | **-0.179** |
| LEU | 0.022 | 0.026 | -0.030 | -0.016 | -0.073 | -0.072 |
| LYS | -0.143 | -0.090 | **0.229** | **0.202** | **-0.534** | **-0.429** |
| MET | 0.083 | 0.057 | -0.009 | 0.006 | -0.135 | -0.092 |
| PHE | -0.043 | -0.025 | -0.005 | 0.003 | -0.137 | -0.058 |
| PRO | 0.028 | -0.012 | -0.092 | -0.041 | 0.052 | 0.006 |
| SER | -0.033 | 0.001 | -0.144 | -0.108 | -0.066 | -0.008 |
| THR | 0.015 | 0.005 | -0.111 | -0.056 | 0.115 | 0.098 |
| TRP | 0.138 | 0.050 | -0.181 | -0.158 | -0.003 | -0.074 |
| TYR | -0.064 | -0.048 | 0.123 | 0.077 | -0.104 | -0.061 |
| VAL | 0.003 | -0.030 | -0.013 | -0.026 | 0.123 | 0.073 |

**Figure 5**
**Species-specific genome and fold composition scoring functions** Species-specific genome (CG) and fold (CF) composition scoring functions derived from the amino acid composition of all predicted ORFs or modeled fold regions, respectively, from the complete genomes of *E. coli, M. jannaschii* and *Halobacterium*. See text for reference to values in bold.

### *Composition-based scoring unctions*

Since there exists significant amino acid variability between protein sequences from different organisms, we sought to generate a scoring function that would allow species-specific identification of protein sequences. Two scoring functions indicating the log probability of amino acid occurrence were generated for each organism. The first scoring function, $C_G$, is based on genomic composition and was derived by taking the log of the amino acid frequency across all genomic ORFs for the given organism over the average amino acid frequency of all the genomes. The second scoring function, $C_F$, was generated from fold composition of the aligned sequences and was derived by taking the log of the amino acid frequencies from the aligned residues of the genomic sequence divided by the template residues. In this fashion the reference state for these scoring functions is what we have termed the 'random organism' since it represents a collection of amino acid compositions from a variety of organisms. This then provides the noise of the scoring function from which we are trying to extract a meaningful, species-specific signal. Log-odds potentials of protein substructures are considered additive [49], and in the evaluation of a sequence, the overall score for a sequence is calculated from the sum of the species-specific log-odds scores for each of its residues.

The nature of these scoring functions is such that if the composition of the organism is not particularly different than the 'random organism', then the magnitude of the scoring function values will approach 0. For instance, the magnitude of the Ecoli $C_G$ and $C_F$ scoring functions values are typically less than either the Mjan or Halob (Figure 5). The $C_G$ and $C_F$ scoring functions are fairly similar and correlate well (86 ± 13%) with each other across all genomes. Mjan has a strong preference for Ile and Lys, but not Gln or Ala largely due to the amino acid coding due to the GC content of the genome (see PCA section). In contrast, Halob prefers the small hydrophobic Ala residue and the charged Asp residue, but not the amidated Asn nor the positively charged Lys. Thus, these scoring functions reflect the probability of observing any residue in a protein sequence or fold for some genome and are heavily influenced by the GC content of the genome and its residue-based environmental adaptations.

### *Cross-validation*

As a preliminary test, we evaluated the performance of the $C_F$ scoring functions for their ability to detect folds in a species-specific manner. That is, the successful scoring function should identify fold sequences of the parent taxonomy from which the scoring function was derived. The performance of the scoring function was evaluated via a jackknife method in which 10% of the model-template pairs were excluded in generating the scoring function. These excluded pairs were then scored with the exclusive scoring function and success was achieved when the score obtained from the model fold was greater than the template fold. The binary species detection ability of the $C_F$ scoring functions to select between the model over the template ranged from 65% to 99% with an average of 85 ± 8% of model sequences being detected from the species-specific fold database (random = 50%). The best $C_F$ detections (>95%) were made with scoring functions derived from those organism found to vary the most in composition including Mpul (99.4%), Buch, Bbur, Halob, Hpylo, Mjan, Mgen, Uure (96.2%). In contrast, the poorest $C_F$ detections were made by common bacteria and pathogen scoring functions from Ecoli variants, Cele, Hsap, Nmeni and Sent. The poor results from these scoring functions reflect the similar model-template composition. In fact, the Ecoli variants obtained ~50% of their template structures from *E. coli*, Cele obtained ~40% of template structures from human, Hsap obtained ~25% of its structures from mouse and 15% from rat and Mmus obtained 46% of template structures from human. The exclusion, or at least the limit of these structure templates would increase the difference in model-template composition and likely generate a more useful scoring function. Thus, these results indicate the admirable species-specific detection ability of the $C_F$ scoring functions on short species-specific domain sequences. Cross-validation was not performed for the $C_G$ scoring functions.

### *Prediction set*

We used all 100 $C_G$ and 95 $C_F$ scoring functions to score every predicted protein sequence from all complete genomes in order to evaluate species-specificity (see Figure 9 (Table 1). The purpose of this experiment is to evaluate the scoring function effectiveness in identifying proteins from the parent organism. Log odd scores were obtained for each protein from each of the complete genomes as evaluated by each of the scoring functions. We also recorded the overall average score obtained by each scoring function across all the ORFs in the genome. In doing so, we discovered that the self-scoring function invariably obtained the lowest overall score (data not shown). The random probability that a scoring function will obtain the best score is determined by the number of best scores included over the total number of scoring functions (i.e. for $C_G$ 10/100 for the top 10 scoring functions using a total of 100 scoring functions) and we can find the maximum value as the difference between the observed success rate and the random probability (Figure 6). We find that the maximum success rate occurs when >20 $C_G$ scoring results are considered. However, as a more conservative estimate, one may choose to consider at least the top 5–10 scoring results to overcome the fact that similar scoring functions obtained by effectively redundant genomes will split the number of successful detections. For instance, scoring functions derived from *E. coli* strains and compositionally similar species (Sent and Styp and their variants) obtained comparable scores, which prevented effective detection of *E. coli* sequences by any of *E. coli* scoring functions when only the top score was considered a detection success. The effect of increasing the number of best scores included from 1 to 5, 10 and 15 can be seen for all scoring functions in Figure 7. The ability of the $C_G$ scoring functions to identify proteins from the parent organism when considering the top 15 scoring results ranged from 51% (EcoliE) to 87% (Paby) with an average 73 ± 9% success. The most effective scoring functions were derived from the low GC organism (Wbre, Buch, Bbur, Baphi, Mpul), hyperthermophiles, Halob and several high GC organisms (Ccres, Mtub, Mtub, Scoel, Smel). When including the top 5 scoring results, the success rate decreased to 49 ± 17%. Note the success rate is significantly higher than random (15/100 or 15%, 5/100 or 5%). In contrast, the effectiveness of the $C_F$ scoring functions varied more

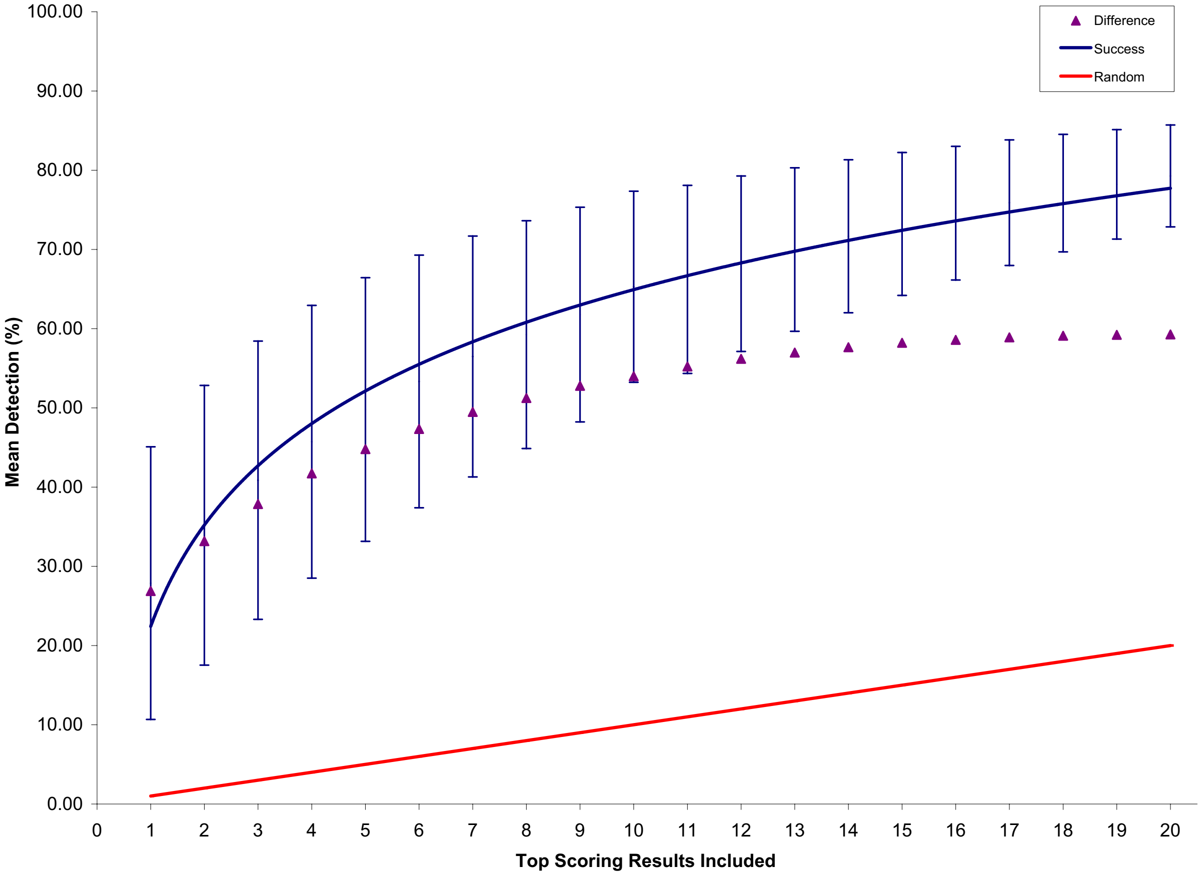


**Figure 6**
**$C_G$ Increased detection when including up to 20 top scoring results** The average success rate determined for scoring functions detecting sequences from their parent organism. The average success rate increases as a logarithmic function while increasing the number of top scoring results (blue). The random probability that a scoring function will detect the sequence is a linear function (red). The maximum difference between the observed success and the random probability occurs when 15 or 16 top scores are included for successful detection. Error bars included for average success.

across this dataset, ranging from a low of 2% (Cele) to a high of 92% (Mpul) with an average success rate of 55 ± 24% when using the top 15 scores, which decreases to 40 ± 25% when only including the top 5. The most successful scoring functions were derived primarily from GC or AT rich organisms. Taken together, the most successful composition-based scoring functions were those derived by organisms with significant composition bias either as a result of %GC skew or from a more extreme environmental niche such as is the case for hyperthermophiles, thermophiles and halophiles. Finally, these results indicate that amino acid composition-based scoring functions may be able to identify the taxonomic origin of protein sequences.

## Conclusions

In the largest study of its kind, we have identified species-specific amino acid composition differences across the predicted open-reading frames of 100 complete genomes. Continuously updated results are available at [http://genome.mshri.on.ca]. Our principal components analysis supports the idea that environmental niche is a major factor for the amino acid composition differences found between species. However, our results raise the possibility that this principal component corresponds more to a complex mixture of environmental influences such as pH, pressure, salt and solute concentrations and to some lesser extent, growth temperature [8,26].

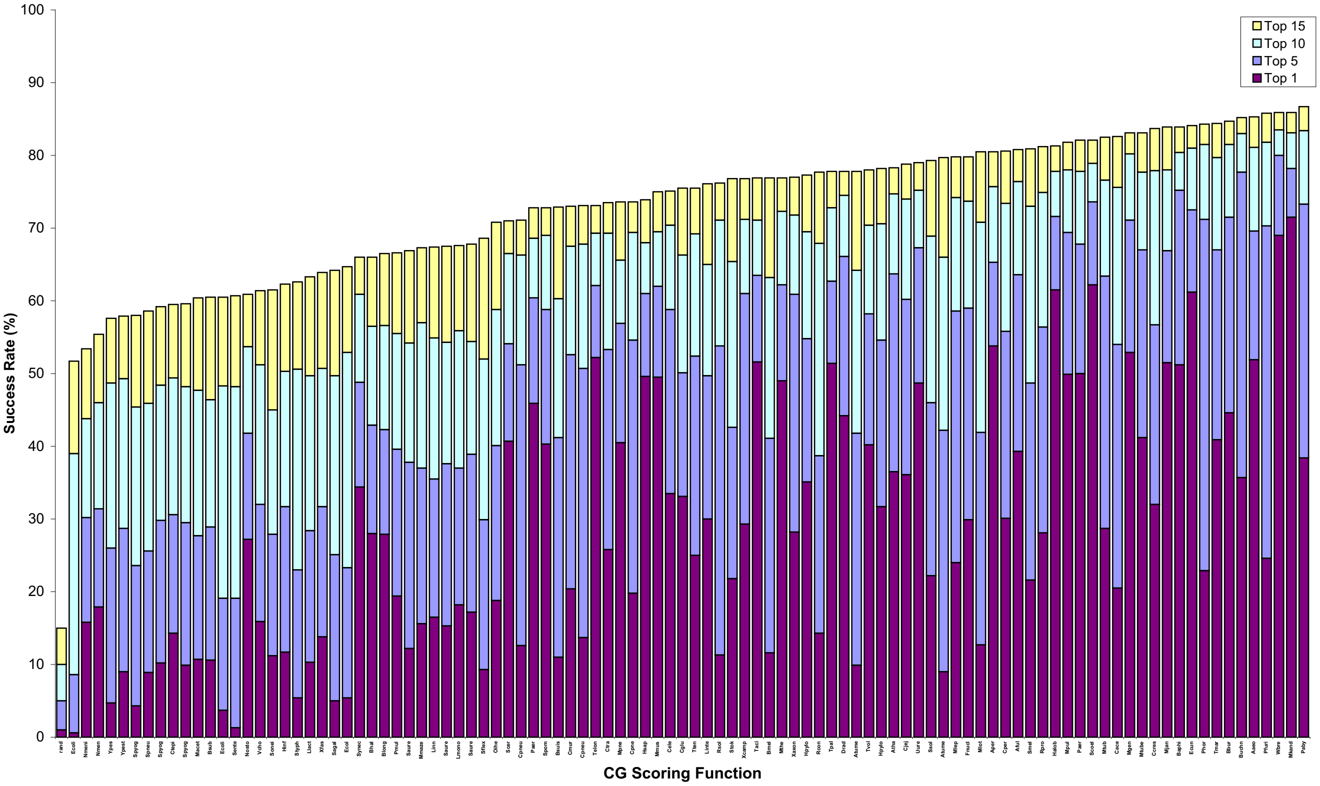


**Figure 7**
**Effect of increasing number of top scores included for detection success with 100 $C_G$ scoring functions** Detection rate increases for increased the number of included scores. Note that certain scoring functions naturally have a high success rate when just considering the top score (Halo, Ecun, MkanA), but others (EcolE, EcolO, Ecol) are redundant and do not necessarily obtain the top score. The former change little when including the top 5, 10 or 15 scores, but the latter largely benefit from this inclusion.

We observed an increased preference for small hydrophobic and charged residue over larger aromatic residues across all species after conservatively modeling 57,840 folds. Moreover, these fold composition biases also illustrate species-specific residue preferences. These biases provided an opportunity for the first time to derive and test simple yet effective species-specific scoring functions. We found that the fold scoring functions are 85 ± 8% effective in detecting a species-specific fold sequence. Moreover, we found that the genomic composition scoring function successfully identified sequences from its parent organism with a surprising 73 ± 9% overall accuracy.

The species-specific composition bias suggests that the variable amino acids are available for structural and/or environmental optimization aspects of proteins. We are currently investigating the usefulness of the species-specific composition-based scoring functions in identifying variable composition regions of protein structures and whether they correspond to structural/functional regions. We are also investigating the possibility of using these scoring functions to find proteins that are non-native to an organism, possibly indicating horizontal transfer. Scoring functions derived from this work can be used in future species-specific protein and fold identification and sequence optimization experiments.

## Methods

Non-redundant protein sequences determined from each of the complete genomes (Table 1) was obtained from the National Center for Biotechnology Information (NCBI – [http://www.ncbi.nlm.nih.gov]) [50] via SeqHound, our integrated sequence and structure database manager [51] [http://seqhound.mshri.on.ca]. Amino acid compositions were computed using all protein coding regions for each complete genome by software developed in our laboratory. Principal components analysis and the amino acid composition dendrogram were generated using the S-PLUS statistics package. Two-tailed paired t-tests were performed to test the null hypothesis that the ORF vs fold mean compositions for each amino acid were the same. All applications were written in ANSI C using the cross-

```
A
  Domain 1 alignment:
  SEQ:    MKHLISMKDIGKEEILEILDEARKMEELLNTKRPLKLLEGKILATVFYEPSTRTRLSFETAMKRLGGEVITMTDLKSSSVAKGESLIDTIRVISGYAD
  STR:    MKHLTTMSELSTEEIKDLLQTAQELKSGKTDNQ----LTGKFAANLFFEPSTRTRFSFEVAEKKLGMNVLNL-DGTSTSVQKGETLYDTIRTLESIGV
  SS:     .bBBBb....aaaAAAAAAAAAAAaaa......----..bbBBBBBbb...aaaAAAAAAaaa..bBBBBBB-.............aaaAAAaaa.bb

  MERGE:  mkhlISmKDiGKeeiLEIlDEaRKMEELLNTKRXXXXlEgkILaTVfYepstrtrLsfeTaMkRlgGEvITMXdLKsTsvAkgeSlIdtirVISGYAD
  COMP:   mkhlISmKDiGKeeiLEIlDEaRKMEELLNTKRlEgkILaTVfYepstrtrLsfeTaMkRlgGEvITMdLKTsvAkgeSlIdtirVIsGYAD

  41% Identity, 100% Occupancy

  Threshold:  25% Identity, 75% Occupancy over structural domain
☑ FINAL:  mkhlISmKDiGKeeiLEIlDEaRKMEELLNTKRlEgkILaTVfYepstrtrLsfeTaMkRlgGEvITMdLKTsvAkgeSlIdtirVIsGYAdII

B
  Domain 2 alignment:
  SEQ:    NVEMYFVSPKELRLPKDIIEDLKAKNIKFYEKESLDDLDDDIDVLYVTRIQKERFPDPNEYEKVKGSYKIKREYVEGKK--FIIMHPLPRVDEIDYD
  STR:    ARVLFSGPS-----------EWQDEENTFGTYVSMDEAVESSDVVMLLRIQNERHQSAVSQEGYLNKYGLTVERAERMKRHAIIMHPAPVNRGVEID
  SS:     bBBBbbb..-----------........bbbbbb........bBBBBBbb...........aaaaaa..............bBBBBb..........

  MERGE:  NVEMYFVSPXXXXXXXXXXXDLKAKNIKfYEKEsLdDLDDDIdvLYVTriqKerFPDPNEYeKVKGSyKIKReYVeGKkXXFiimhpLpRVDEIDYd
  COMP:   NVEMYFVSPDLKAKNIKfYEKEsLdDLDDDIdvLYVTriqKerFPDPNEYeKVKGSyKIKReYVeGKkXXFiimhpLpRVDEIDYD

  20% Identity, 97% Occupancy

  Threshold:  25% Identity, 75% Occupancy over structural domain
⊘ FINAL:  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

**Figure 8**
**Sequence to structural domain alignment** Sequence to structural domain alignments (A, B). A genomic sequence (SEQ) is aligned to a homologous sequence with a 3D structure (STR) using a secondary structure profile using ClustalW. Note the insertion of gaps (denoted by -, red) in non-structured regions of the 3D structure. In the MERGE step, gaps in the structure are masked out, and eliminated in the compression step (COMP). At this point, the number of identical residues and the number of residues in the genomic sequence occupying a domain position in the structure are counted. Since domain 1 alignment passes the minimal 25% identity and 75% occupancy, it is used for further analysis. However, the %identity in the domain 2 alignment (B) is lower than the threshold of 25%, and the entire domain alignment is masked out and not used in any further analyses.

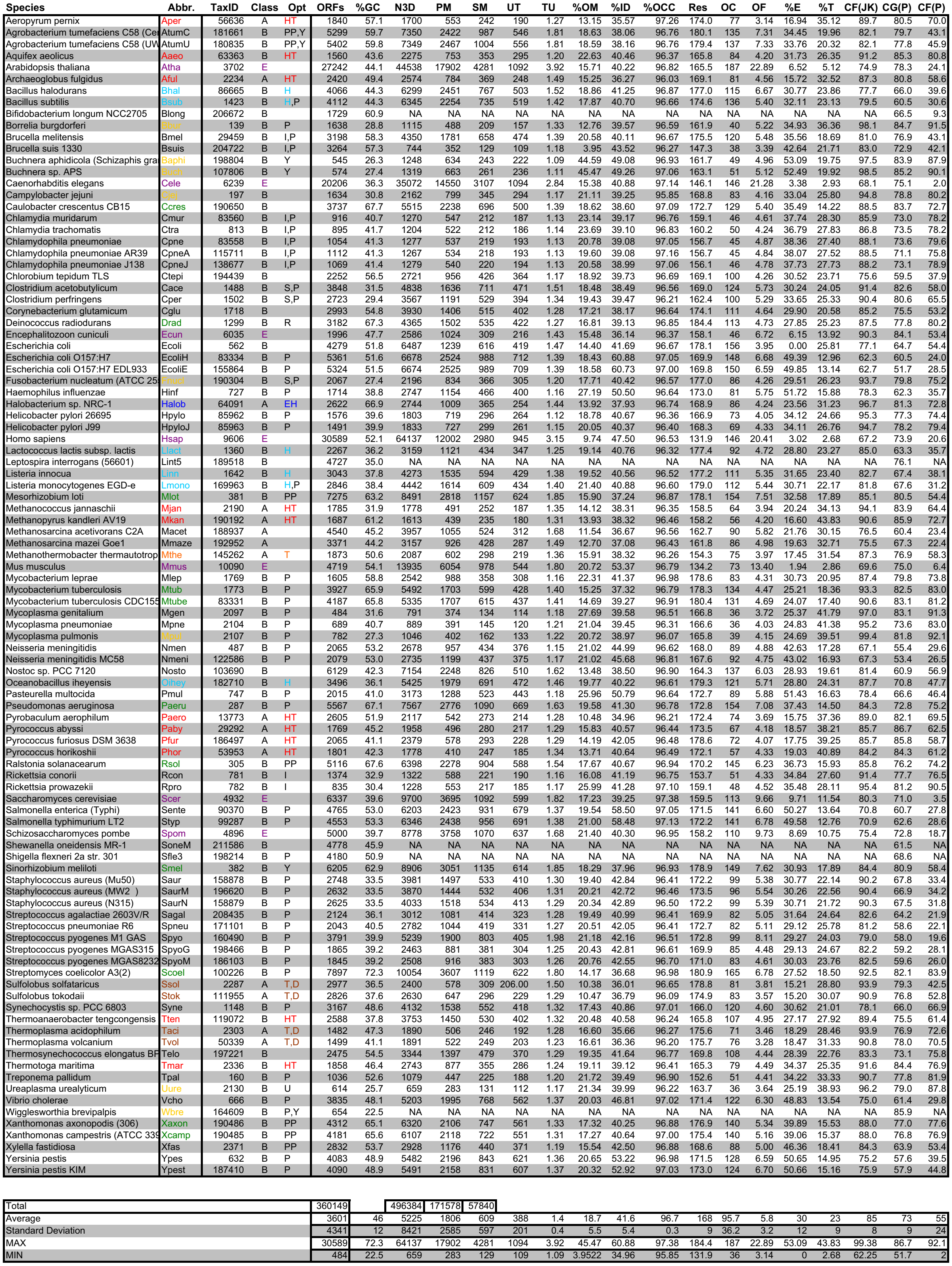
platform NCBI Toolkit [http://www.ncbi.nlm.nih.gov/IEB] and have been compiled and tested on Windows 98/ME/NT/2000/XP, MacOsX, Linux, HP-UX, PA-RISC Linux, Compaq Tru64, IRIX, Solaris, QNX, FreeBSD and PowerPC-Linux operating systems. Protein sequence and fold scoring functions are available as additional files.

### *Conservative fold modeling*

Domains are the fundamental unit of a polypeptide chain or part of a polypeptide chain that are thought to independently fold into a stable tertiary structure. Since domains are often units of function and different domains of a protein are often associated with different functions, we evaluated sequence alignments on a structural domain-by-domain basis rather than by the global alignment. This provides a conservative framework to evaluate structural alignments.

#### *Sequence and structure*

For each protein sequence from a completely sequenced and annotated genome, herein referred to as a genomic sequence, we identified neighbour sequences, that is, sequences in the non-redundant protein sequence database sharing significant levels of similarity (expect value < 0.01) using NBLAST, a cluster-computer variant of BLAST [52] (Table 1). No efforts were made to minimize a possible bias contributed by paralogous genomic sequences. Neighbour sequences with 3D structures, herein referred to as templates, were identified using SeqHound [51], in a similar fashion to the NCBI's genome annotation service [53]. These genomic sequences and their corresponding templates are then used to generate hi-fidelity sequence to structure alignments.



| Species | Abbr. | TaxID | Class | Opt | ORFs | %GC | N3D | PM | SM | UT | TU | %OM | %ID | %OCC | Res | OC | OF | %E | %T | CF(JK) | CG(P) | CF(P) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aeropyrum pernix | Aper | 56636 | A | HT | 1840 | 57.1 | 1700 | 553 | 242 | 190 | 1.27 | 13.15 | 35.57 | 97.26 | 174.0 | 77 | 3.14 | 16.94 | 35.12 | 89.7 | 80.5 | 70.0 |
| Agrobacterium tumefaciens C58 (Ce | AtumC | 181661 | B | PP,Y | 5299 | 59.7 | 7350 | 2422 | 987 | 546 | 1.81 | 18.63 | 38.06 | 96.76 | 180.1 | 135 | 7.31 | 34.45 | 19.96 | 82.1 | 79.7 | 43.1 |
| Agrobacterium tumefaciens C58 (UW | AtumU | 180835 | B | PP,Y | 5402 | 59.8 | 7349 | 2467 | 1004 | 556 | 1.81 | 18.59 | 38.16 | 96.76 | 179.4 | 137 | 7.33 | 33.76 | 20.32 | 82.1 | 77.8 | 45.9 |
| Aquifex aeolicus | Aaeo | 63363 | B | HT | 1560 | 43.6 | 2275 | 753 | 353 | 295 | 1.20 | 22.63 | 40.46 | 96.37 | 165.8 | 84 | 4.20 | 31.73 | 26.35 | 91.2 | 85.3 | 80.8 |
| Arabidopsis thaliana | Atha | 3702 | E |  | 27242 | 44.1 | 44538 | 17902 | 4281 | 1092 | 3.92 | 15.71 | 40.22 | 96.82 | 165.5 | 187 | 22.89 | 6.52 | 5.12 | 74.9 | 78.3 | 24.1 |
| Archaeoglobus fulgidus | Aful | 2234 | A | HT | 2420 | 49.4 | 2574 | 784 | 369 | 248 | 1.49 | 15.25 | 36.27 | 96.03 | 169.1 | 81 | 4.56 | 15.72 | 32.52 | 87.3 | 80.8 | 58.6 |
| Bacillus halodurans | Bhal | 86665 | B | H | 4066 | 44.3 | 6299 | 2451 | 767 | 503 | 1.52 | 18.86 | 41.25 | 96.87 | 177.0 | 115 | 6.67 | 30.77 | 23.86 | 77.7 | 66.0 | 39.6 |
| Bacillus subtilis | Bsub | 1423 | B | H,P | 4112 | 44.3 | 6345 | 2254 | 735 | 519 | 1.42 | 17.87 | 40.70 | 96.66 | 174.6 | 136 | 5.40 | 32.11 | 23.13 | 79.5 | 60.5 | 30.6 |
| Bifidobacterium longum NCC2705 | Blong | 206672 | B |  | 1729 | 60.9 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 66.5 | 9.3 |
| Borrelia burgdorferi | Bbur | 139 | B | P | 1638 | 28.8 | 1115 | 488 | 209 | 157 | 1.33 | 12.76 | 39.57 | 96.59 | 161.9 | 40 | 5.22 | 34.93 | 36.36 | 98.1 | 84.7 | 91.5 |
| Brucella melitensis | Bmel | 29459 | B | I,P | 3198 | 58.3 | 4350 | 1781 | 658 | 474 | 1.39 | 20.58 | 40.11 | 96.67 | 175.5 | 120 | 5.48 | 35.56 | 18.69 | 81.0 | 76.9 | 43.1 |
| Brucella suis 1330 | Bsuis | 204722 | B | I,P | 3264 | 57.3 | 744 | 352 | 129 | 109 | 1.18 | 3.95 | 43.52 | 96.27 | 147.3 | 38 | 3.39 | 42.64 | 21.71 | 83.0 | 72.9 | 42.1 |
| Buchnera aphidicola (Schizaphis gra | Baphi | 198804 | B | Y | 545 | 26.3 | 1248 | 634 | 243 | 222 | 1.09 | 44.59 | 49.08 | 96.93 | 161.7 | 49 | 4.96 | 53.09 | 19.75 | 97.5 | 83.9 | 87.9 |
| Buchnera sp. APS | Buch | 107806 | B | Y | 574 | 27.4 | 1319 | 663 | 261 | 236 | 1.11 | 45.47 | 49.26 | 97.06 | 163.1 | 51 | 5.12 | 52.49 | 19.92 | 98.5 | 85.2 | 90.1 |
| Caenorhabditis elegans | Cele | 6239 | E |  | 20206 | 36.3 | 35072 | 14550 | 3107 | 1094 | 2.84 | 15.38 | 40.88 | 97.14 | 146.1 | 146 | 21.28 | 3.38 | 2.93 | 68.1 | 75.1 | 2.0 |
| Campylobacter jejuni | Cjej | 197 | B |  | 1634 | 30.8 | 2162 | 799 | 345 | 294 | 1.17 | 21.11 | 39.25 | 95.85 | 168.8 | 83 | 4.16 | 33.04 | 25.80 | 94.8 | 78.8 | 80.2 |
| Caulobacter crescentus CB15 | Ccres | 190650 | B |  | 3737 | 67.7 | 5515 | 2238 | 696 | 500 | 1.39 | 18.62 | 38.60 | 97.09 | 172.7 | 129 | 5.40 | 35.49 | 14.22 | 88.5 | 83.7 | 72.7 |
| Chlamydia muridarum | Cmur | 83560 | B | I,P | 916 | 40.7 | 1270 | 547 | 212 | 187 | 1.13 | 23.14 | 39.17 | 96.76 | 159.1 | 46 | 4.61 | 37.74 | 28.30 | 85.9 | 73.0 | 78.2 |
| Chlamydia trachomatis | Ctra | 813 | B | I,P | 895 | 41.7 | 1204 | 522 | 212 | 186 | 1.14 | 23.69 | 39.10 | 96.83 | 160.2 | 50 | 4.24 | 36.79 | 27.83 | 86.8 | 73.5 | 78.2 |
| Chlamydophila pneumoniae | Cpne | 83558 | B | I,P | 1054 | 41.3 | 1277 | 537 | 219 | 193 | 1.13 | 20.78 | 39.08 | 97.05 | 156.7 | 45 | 4.87 | 38.36 | 27.40 | 88.1 | 73.6 | 79.6 |
| Chlamydophila pneumoniae AR39 | CpneA | 115711 | B | I,P | 1112 | 41.3 | 1267 | 534 | 218 | 193 | 1.13 | 19.60 | 39.08 | 97.16 | 156.7 | 45 | 4.84 | 38.07 | 27.52 | 88.5 | 71.1 | 75.8 |
| Chlamydophila pneumoniae J138 | CpneJ | 138677 | B | I,P | 1069 | 41.4 | 1279 | 540 | 220 | 194 | 1.13 | 20.58 | 38.99 | 97.06 | 156.1 | 46 | 4.78 | 37.73 | 27.73 | 88.2 | 73.1 | 78.9 |
| Chlorobium tepidum TLS | Ctepi | 194439 | B |  | 2252 | 56.5 | 2721 | 956 | 426 | 364 | 1.17 | 18.92 | 39.73 | 96.69 | 169.1 | 100 | 4.26 | 30.52 | 23.71 | 75.6 | 59.5 | 37.9 |
| Clostridium acetobutylicum | Cace | 1488 | B | S,P | 3848 | 31.5 | 4838 | 1636 | 711 | 471 | 1.51 | 18.48 | 38.49 | 96.56 | 169.0 | 124 | 5.73 | 30.24 | 24.05 | 91.4 | 82.6 | 58.0 |
| Clostridium perfringens | Cper | 1502 | B | S,P | 2723 | 29.4 | 3567 | 1191 | 529 | 394 | 1.34 | 19.43 | 39.47 | 96.21 | 162.4 | 100 | 5.29 | 33.65 | 25.33 | 90.4 | 80.6 | 65.5 |
| Corynebacterium glutamicum | Cglu | 1718 | B |  | 2993 | 54.8 | 3930 | 1406 | 515 | 402 | 1.28 | 17.21 | 38.17 | 96.64 | 174.1 | 111 | 4.64 | 29.90 | 20.58 | 85.2 | 75.5 | 53.2 |
| Deinococcus radiodurans | Drad | 1299 | B | R | 3182 | 67.3 | 4365 | 1502 | 535 | 422 | 1.27 | 16.81 | 39.13 | 96.85 | 184.4 | 113 | 4.73 | 27.85 | 25.23 | 87.5 | 77.8 | 80.2 |
| Encephalitozoon cuniculi | Ecun | 6035 | E |  | 1996 | 47.7 | 2586 | 1024 | 309 | 216 | 1.43 | 15.48 | 36.14 | 96.37 | 158.1 | 46 | 6.72 | 6.15 | 13.92 | 90.3 | 84.1 | 53.4 |
| Escherichia coli | Ecoli | 562 | B |  | 4279 | 51.8 | 6487 | 1239 | 616 | 419 | 1.47 | 14.40 | 41.69 | 96.67 | 178.1 | 156 | 3.95 | 0.00 | 25.81 | 77.1 | 64.7 | 54.4 |
| Escherichia coli O157:H7 | EcoliH | 83334 | B | P | 5361 | 51.6 | 6678 | 2524 | 988 | 712 | 1.39 | 18.43 | 60.88 | 97.05 | 169.9 | 148 | 6.68 | 49.39 | 12.96 | 62.3 | 60.5 | 24.0 |
| Escherichia coli O157:H7 EDL933 | EcoliE | 155864 | B | P | 5324 | 51.5 | 6674 | 2525 | 989 | 709 | 1.39 | 18.58 | 60.73 | 97.00 | 169.8 | 150 | 6.59 | 49.85 | 13.14 | 62.7 | 51.7 | 28.5 |
| Fusobacterium nucleatum (ATCC 25 | Fnucl | 190304 | B | S,P | 2067 | 27.4 | 2196 | 834 | 366 | 305 | 1.20 | 17.71 | 40.42 | 96.57 | 177.0 | 86 | 4.26 | 29.51 | 26.23 | 93.7 | 79.8 | 75.2 |
| Haemophilus influenzae | Hinf | 727 | B | P | 1714 | 38.8 | 2747 | 1154 | 466 | 400 | 1.16 | 27.19 | 50.50 | 96.64 | 173.0 | 81 | 5.75 | 51.72 | 15.88 | 78.3 | 62.3 | 35.7 |
| Halobacterium sp. NRC-1 | Halob | 64091 | A | EH | 2622 | 66.9 | 2744 | 1009 | 365 | 254 | 1.44 | 13.92 | 37.93 | 96.74 | 168.9 | 86 | 4.24 | 23.56 | 31.23 | 96.7 | 81.3 | 72.8 |
| Helicobacter pylori 26695 | Hpylo | 85962 | B | P | 1576 | 39.6 | 1803 | 719 | 296 | 264 | 1.12 | 18.78 | 40.67 | 96.36 | 166.9 | 73 | 4.05 | 34.12 | 24.66 | 95.3 | 77.3 | 74.4 |
| Helicobacter pylori J99 | HpyloJ | 85963 | B | P | 1491 | 39.9 | 1833 | 727 | 299 | 261 | 1.15 | 20.05 | 40.37 | 96.40 | 168.3 | 69 | 4.33 | 34.11 | 26.76 | 94.7 | 78.2 | 79.4 |
| Homo sapiens | Hsap | 9606 | E |  | 30589 | 52.1 | 64137 | 12002 | 2980 | 945 | 3.15 | 9.74 | 47.50 | 96.53 | 131.9 | 146 | 20.41 | 3.02 | 2.68 | 67.2 | 73.9 | 20.6 |
| Lactococcus lactis subsp. lactis | Llact | 1360 | B | H | 2267 | 36.2 | 3159 | 1121 | 434 | 347 | 1.25 | 19.14 | 40.76 | 96.32 | 177.4 | 92 | 4.72 | 28.80 | 23.27 | 85.0 | 63.3 | 35.7 |
| Leptospira interrogans (56601) | Lint5 | 189518 | B |  | 4727 | 35.0 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 76.1 | NA |
| Listeria innocua | Linn | 1642 | B | H | 3043 | 37.8 | 4273 | 1535 | 594 | 429 | 1.38 | 19.52 | 40.56 | 96.52 | 177.2 | 111 | 5.35 | 31.65 | 23.40 | 82.7 | 67.4 | 38.1 |
| Listeria monocytogenes EGD-e | Lmono | 169963 | B | H,P | 2846 | 38.4 | 4442 | 1614 | 609 | 434 | 1.40 | 21.40 | 40.88 | 96.60 | 179.0 | 112 | 5.44 | 30.71 | 22.17 | 81.8 | 67.6 | 31.2 |
| Mesorhizobium loti | Mlot | 381 | B | PP | 7275 | 63.2 | 8491 | 2818 | 1157 | 624 | 1.85 | 15.90 | 37.24 | 96.87 | 178.1 | 154 | 7.51 | 32.58 | 17.89 | 85.1 | 80.5 | 54.4 |
| Methanococcus jannaschii | Mjan | 2190 | A | HT | 1785 | 31.9 | 1778 | 491 | 252 | 187 | 1.35 | 14.12 | 38.31 | 96.35 | 158.5 | 64 | 3.94 | 20.24 | 34.13 | 94.1 | 83.9 | 64.4 |
| Methanopyrus kandleri AV19 | Mkan | 190192 | A | HT | 1687 | 61.2 | 1613 | 439 | 235 | 180 | 1.31 | 13.93 | 38.32 | 96.46 | 158.2 | 56 | 4.20 | 16.60 | 43.83 | 90.6 | 85.9 | 72.7 |
| Methanosarcina acetivorans C2A | Macet | 188937 | A |  | 4540 | 45.2 | 3957 | 1055 | 524 | 312 | 1.68 | 11.54 | 36.67 | 96.56 | 162.7 | 90 | 5.82 | 21.76 | 30.15 | 76.5 | 60.4 | 23.4 |
| Methanosarcina mazei Goe1 | Mmaze | 192952 | A |  | 3371 | 44.2 | 3157 | 926 | 428 | 287 | 1.49 | 12.70 | 37.08 | 96.43 | 161.8 | 86 | 4.98 | 19.63 | 32.71 | 75.5 | 67.3 | 22.4 |
| Methanothermobacter thermautotrop | Mthe | 145262 | A | T | 1873 | 50.6 | 2087 | 602 | 298 | 219 | 1.36 | 15.91 | 38.32 | 96.26 | 154.3 | 75 | 3.97 | 17.45 | 31.54 | 87.3 | 76.9 | 58.3 |
| Mus musculus | Mmus | 10090 | E |  | 4719 | 54.1 | 13935 | 6054 | 978 | 544 | 1.80 | 20.72 | 53.37 | 96.79 | 134.2 | 73 | 13.40 | 1.94 | 2.86 | 69.6 | 75.0 | 6.4 |
| Mycobacterium leprae | Mlep | 1769 | B | P | 1605 | 58.8 | 2542 | 988 | 358 | 308 | 1.16 | 22.31 | 41.37 | 96.98 | 178.6 | 83 | 4.31 | 30.73 | 20.95 | 87.4 | 79.8 | 73.8 |
| Mycobacterium tuberculosis | Mtub | 1773 | B | P | 3927 | 65.9 | 5492 | 1703 | 599 | 428 | 1.40 | 15.25 | 37.32 | 96.79 | 178.3 | 134 | 4.47 | 25.21 | 18.36 | 93.3 | 82.5 | 83.0 |
| Mycobacterium tuberculosis CDC155 | Mtube | 83331 | B | P | 4187 | 65.8 | 5335 | 1707 | 615 | 437 | 1.41 | 14.69 | 39.27 | 96.91 | 180.4 | 131 | 4.69 | 24.07 | 17.40 | 90.6 | 83.1 | 81.2 |
| Mycoplasma genitalium | Mgen | 2097 | B | P | 484 | 31.6 | 791 | 374 | 134 | 114 | 1.18 | 27.69 | 39.58 | 96.51 | 166.8 | 36 | 3.72 | 25.37 | 41.79 | 97.0 | 83.1 | 91.3 |
| Mycoplasma pneumoniae | Mpne | 2104 | B | P | 689 | 40.7 | 889 | 391 | 145 | 120 | 1.21 | 21.04 | 39.45 | 96.31 | 166.6 | 36 | 4.03 | 24.83 | 41.38 | 95.2 | 73.6 | 83.0 |
| Mycoplasma pulmonis | Mpul | 2107 | B | P | 782 | 27.3 | 1046 | 402 | 162 | 133 | 1.22 | 20.72 | 38.97 | 96.07 | 165.8 | 39 | 4.15 | 24.69 | 39.51 | 99.4 | 81.8 | 92.1 |
| Neisseria meningitidis | Nmen | 487 | B | P | 2065 | 53.2 | 2678 | 957 | 434 | 376 | 1.15 | 21.02 | 44.99 | 96.62 | 168.0 | 89 | 4.88 | 42.63 | 17.28 | 67.1 | 55.4 | 29.6 |
| Neisseria meningitidis MC58 | Nmeni | 122586 | B | P | 2079 | 53.0 | 2735 | 1199 | 437 | 375 | 1.17 | 21.02 | 45.68 | 96.81 | 167.6 | 92 | 4.75 | 43.02 | 16.93 | 67.3 | 53.4 | 26.5 |
| Nostoc sp. PCC 7120 | Nosto | 103690 | B |  | 6129 | 42.3 | 7154 | 2248 | 826 | 510 | 1.62 | 13.48 | 38.50 | 96.90 | 164.3 | 137 | 6.03 | 28.93 | 19.61 | 81.4 | 60.9 | 56.9 |
| Oceanobacillus iheyensis | Oihey | 182710 | B | H | 3496 | 36.1 | 5425 | 1979 | 691 | 472 | 1.46 | 19.77 | 40.22 | 96.61 | 179.3 | 121 | 5.71 | 28.80 | 24.31 | 87.7 | 70.8 | 47.7 |
| Pasteurella multocida | Pmul | 747 | B | P | 2015 | 41.0 | 3173 | 1288 | 523 | 443 | 1.18 | 25.96 | 50.79 | 96.64 | 172.7 | 89 | 5.88 | 51.43 | 16.63 | 78.4 | 66.6 | 46.4 |
| Pseudomonas aeruginosa | Paeru | 287 | B | P | 5567 | 67.1 | 7567 | 2776 | 1090 | 669 | 1.63 | 19.58 | 41.30 | 96.78 | 172.8 | 154 | 7.08 | 37.43 | 14.50 | 84.3 | 72.8 | 75.2 |
| Pyrobaculum aerophilum | Paero | 13773 | A | HT | 2605 | 51.9 | 2117 | 542 | 273 | 214 | 1.28 | 10.48 | 34.96 | 96.21 | 172.4 | 74 | 3.69 | 15.75 | 37.36 | 89.0 | 82.1 | 69.5 |
| Pyrococcus abyssi | Paby | 29292 | A | HT | 1769 | 45.2 | 1958 | 496 | 280 | 217 | 1.29 | 15.83 | 40.57 | 96.44 | 173.5 | 67 | 4.18 | 18.57 | 38.21 | 85.7 | 86.7 | 62.5 |
| Pyrococcus furiosus DSM 3638 | Pfur | 186497 | A | HT | 2065 | 41.1 | 2379 | 578 | 293 | 228 | 1.29 | 14.19 | 42.05 | 96.48 | 178.6 | 72 | 4.07 | 17.75 | 39.25 | 85.7 | 85.8 | 58.7 |
| Pyrococcus horikoshii | Phor | 53953 | A | HT | 1801 | 42.3 | 1778 | 410 | 247 | 185 | 1.34 | 13.71 | 40.64 | 96.49 | 172.1 | 57 | 4.33 | 19.03 | 40.89 | 84.2 | 84.3 | 61.2 |
| Ralstonia solanacearum | Rsol | 305 | B | PP | 5116 | 67.6 | 6398 | 2278 | 904 | 588 | 1.54 | 17.67 | 40.67 | 96.94 | 170.2 | 145 | 6.23 | 36.73 | 15.93 | 85.8 | 76.2 | 74.2 |
| Rickettsia conorii | Rcon | 781 | B | I | 1374 | 32.9 | 1322 | 588 | 221 | 190 | 1.16 | 16.08 | 41.19 | 96.75 | 153.7 | 51 | 4.33 | 34.84 | 27.60 | 91.4 | 77.7 | 76.5 |
| Rickettsia prowazekii | Rpro | 782 | B | I | 835 | 30.4 | 1228 | 553 | 217 | 185 | 1.17 | 25.99 | 41.28 | 97.10 | 159.1 | 48 | 4.52 | 35.48 | 28.11 | 95.4 | 81.2 | 90.5 |
| Saccharomyces cerevisiae | Scer | 4932 | E |  | 6337 | 39.6 | 9700 | 3695 | 1092 | 599 | 1.82 | 17.23 | 39.25 | 97.38 | 159.5 | 113 | 9.66 | 9.71 | 11.54 | 80.3 | 71.0 | 3.5 |
| Salmonella enterica (Typhi) | Sente | 90370 | B | P | 4765 | 53.0 | 6203 | 2423 | 931 | 679 | 1.37 | 19.54 | 58.50 | 97.05 | 171.5 | 141 | 6.60 | 50.27 | 13.64 | 70.8 | 60.7 | 27.8 |
| Salmonella typhimurium LT2 | Styp | 99287 | B | P | 4553 | 53.3 | 6346 | 2438 | 956 | 691 | 1.38 | 21.00 | 58.48 | 97.13 | 172.2 | 141 | 6.78 | 49.58 | 12.76 | 70.9 | 62.6 | 28.6 |
| Schizosaccharomyces pombe | Spom | 4896 | E |  | 5000 | 39.7 | 8778 | 3758 | 1070 | 637 | 1.68 | 21.40 | 40.30 | 96.95 | 158.2 | 110 | 9.73 | 8.69 | 10.75 | 75.4 | 72.8 | 18.7 |
| Shewanella oneidensis MR-1 | SoneM | 211586 | B |  | 4778 | 45.9 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 61.5 | NA |
| Shigella flexneri 2a str. 301 | Sfle3 | 198214 | B | P | 4180 | 50.9 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 68.6 | NA |
| Sinorhizobium meliloti | Smel | 382 | B | Y | 6205 | 62.9 | 8906 | 3051 | 1135 | 614 | 1.85 | 18.29 | 37.96 | 96.93 | 178.9 | 149 | 7.62 | 30.93 | 17.89 | 84.4 | 80.9 | 58.4 |
| Staphylococcus aureus (Mu50) | Saur | 158878 | B | P | 2748 | 33.5 | 3981 | 1497 | 533 | 410 | 1.30 | 19.40 | 42.84 | 96.41 | 172.2 | 99 | 5.38 | 30.77 | 22.14 | 90.2 | 67.8 | 33.4 |
| Staphylococcus aureus (MW2 ) | SaurM | 196620 | B | P | 2632 | 33.5 | 3870 | 1444 | 532 | 406 | 1.31 | 20.21 | 42.72 | 96.46 | 173.5 | 96 | 5.54 | 30.26 | 22.56 | 90.4 | 66.9 | 34.2 |
| Staphylococcus aureus (N315) | SaurN | 158879 | B | P | 2625 | 33.5 | 4033 | 1518 | 534 | 413 | 1.29 | 20.34 | 42.89 | 96.50 | 172.2 | 99 | 5.39 | 30.71 | 21.72 | 90.3 | 67.5 | 31.8 |
| Streptococcus agalactiae 2603V/R | Sagal | 208435 | B | P | 2124 | 36.1 | 3012 | 1081 | 414 | 323 | 1.28 | 19.49 | 40.99 | 96.41 | 169.9 | 82 | 5.05 | 31.64 | 24.64 | 82.6 | 64.2 | 21.9 |
| Streptococcus pneumoniae R6 | Spneu | 171101 | B | P | 2043 | 40.5 | 2782 | 1044 | 419 | 331 | 1.27 | 20.51 | 42.05 | 96.41 | 172.7 | 82 | 5.11 | 29.12 | 25.78 | 81.2 | 58.6 | 22.1 |
| Streptococcus pyogenes M1 GAS | Spyo | 160490 | B | P | 3791 | 39.9 | 5239 | 1900 | 803 | 405 | 1.98 | 21.18 | 42.16 | 96.51 | 172.8 | 99 | 8.11 | 29.27 | 24.03 | 79.0 | 58.0 | 19.6 |
| Streptococcus pyogenes MGAS315 | SpyoG | 198466 | B | P | 1865 | 39.2 | 2463 | 881 | 381 | 304 | 1.25 | 20.43 | 42.81 | 96.61 | 169.9 | 85 | 4.48 | 29.13 | 24.67 | 82.2 | 59.2 | 28.1 |
| Streptococcus pyogenes MGAS8232 | SpyoM | 186103 | B | P | 1845 | 39.2 | 2508 | 916 | 383 | 303 | 1.26 | 20.76 | 42.55 | 96.70 | 171.0 | 83 | 4.61 | 30.03 | 23.76 | 82.5 | 59.6 | 26.0 |
| Streptomyces coelicolor A3(2) | Scoel | 100226 | B | P | 7897 | 72.3 | 10054 | 3607 | 1119 | 622 | 1.80 | 14.17 | 36.68 | 96.98 | 180.9 | 165 | 6.78 | 27.52 | 18.50 | 92.5 | 82.1 | 83.9 |
| Sulfolobus solfataricus | Ssol | 2287 | A | T,D | 2977 | 36.5 | 2400 | 578 | 309 | 206.00 | 1.50 | 10.38 | 36.01 | 96.65 | 178.8 | 81 | 3.81 | 15.21 | 28.80 | 93.9 | 79.3 | 42.5 |
| Sulfolobus tokodaii | Stok | 111955 | A | T,D | 2826 | 37.6 | 2630 | 647 | 296 | 229 | 1.29 | 10.47 | 36.79 | 96.09 | 174.9 | 83 | 3.57 | 15.20 | 30.07 | 90.9 | 76.8 | 52.8 |
| Synechocystis sp. PCC 6803 | Syne | 1148 | B | P | 3167 | 48.6 | 4132 | 1538 | 552 | 418 | 1.32 | 17.43 | 40.86 | 97.01 | 166.0 | 120 | 4.60 | 30.62 | 21.01 | 78.1 | 66.0 | 66.9 |
| Thermoanaerobacter tengcongensis | Tten | 119072 | B | HT | 2588 | 37.8 | 3753 | 1450 | 530 | 402 | 1.32 | 20.48 | 40.58 | 96.24 | 165.8 | 107 | 4.95 | 27.17 | 27.92 | 89.4 | 75.5 | 61.4 |
| Thermoplasma acidophilum | Taci | 2303 | A | T,D | 1482 | 47.3 | 1890 | 506 | 246 | 192 | 1.28 | 16.60 | 35.66 | 96.27 | 175.6 | 71 | 3.46 | 18.29 | 28.46 | 93.9 | 76.9 | 72.6 |
| Thermoplasma volcanium | Tvol | 50339 | A | T,D | 1499 | 41.1 | 1891 | 522 | 249 | 203 | 1.23 | 16.61 | 36.36 | 96.20 | 175.7 | 76 | 3.28 | 18.47 | 31.33 | 90.8 | 78.0 | 70.5 |
| Thermosynechococcus elongatus BP | Telo | 197221 | B |  | 2475 | 54.5 | 3344 | 1397 | 479 | 370 | 1.29 | 19.35 | 41.64 | 96.77 | 169.8 | 108 | 4.44 | 28.39 | 22.76 | 83.3 | 73.1 | 75.8 |
| Thermotoga maritima | Tmar | 2336 | B | HT | 1858 | 46.4 | 2743 | 877 | 355 | 286 | 1.24 | 19.11 | 39.12 | 96.41 | 165.3 | 79 | 4.49 | 34.37 | 25.35 | 91.6 | 84.4 | 76.9 |
| Treponema pallidum | Tpal | 160 | B | P | 1036 | 52.6 | 1079 | 447 | 225 | 188 | 1.20 | 21.72 | 39.49 | 96.90 | 152.6 | 51 | 4.41 | 34.22 | 33.33 | 90.7 | 77.8 | 81.9 |
| Ureaplasma urealyticum | Uure | 2130 | B | U | 614 | 25.7 | 659 | 283 | 131 | 112 | 1.17 | 21.34 | 39.99 | 96.22 | 163.7 | 36 | 3.64 | 25.19 | 38.93 | 96.2 | 79.0 | 87.8 |
| Vibrio cholerae | Vcho | 666 | B | P | 3835 | 48.1 | 5203 | 1995 | 768 | 562 | 1.37 | 20.03 | 46.81 | 97.02 | 171.4 | 122 | 6.30 | 48.83 | 13.54 | 75.0 | 61.4 | 29.8 |
| Wigglesworthia brevipalpis | Wbre | 164609 | B | P,Y | 654 | 22.5 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 85.9 | NA |
| Xanthomonas axonopodis (306) | Xaxon | 190486 | B | PP | 4312 | 65.1 | 6320 | 2106 | 747 | 561 | 1.33 | 17.32 | 40.25 | 96.88 | 176.9 | 140 | 5.34 | 39.89 | 15.53 | 88.0 | 77.0 | 77.6 |
| Xanthomonas campestris (ATCC 339 | Xcamp | 190485 | B | PP | 4181 | 65.6 | 6107 | 2118 | 722 | 551 | 1.31 | 17.27 | 40.64 | 97.00 | 175.4 | 140 | 5.16 | 39.06 | 15.37 | 88.0 | 76.8 | 76.9 |
| Xylella fastidiosa | Xfas | 2371 | B | PP | 2832 | 53.7 | 2928 | 1176 | 440 | 371 | 1.19 | 15.54 | 42.50 | 96.88 | 168.6 | 88 | 5.00 | 46.36 | 18.41 | 84.3 | 63.9 | 53.4 |
| Yersinia pestis | Ypes | 632 | B | P | 4083 | 48.9 | 5482 | 2196 | 843 | 621 | 1.36 | 20.65 | 53.22 | 96.98 | 171.5 | 128 | 6.59 | 50.65 | 14.95 | 75.2 | 57.6 | 39.5 |
| Yersinia pestis KIM | Ypest | 187410 | B | P | 4090 | 48.9 | 5491 | 2158 | 831 | 607 | 1.37 | 20.32 | 52.92 | 97.03 | 173.0 | 124 | 6.70 | 50.66 | 15.16 | 75.9 | 57.9 | 44.8 |

|  | ORFs | %GC | N3D | PM | SM | UT | TU | %OM | %ID | %OCC | Res | OC | OF | %E | %T | CF(JK) | CG(P) | CF(P) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | 360149 |  | 496384 | 171578 | 57840 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Average | 3601 | 46 | 5225 | 1806 | 609 | 388 | 1.4 | 18.7 | 41.6 | 96.7 | 168 | 95.7 | 5.8 | 30 | 23 | 85 | 73 | 55 |
| Standard Deviation | 4341 | 12 | 8421 | 2585 | 597 | 201 | 0.4 | 5.5 | 5.4 | 0.3 | 9 | 36.2 | 3.2 | 12 | 9 | 8 | 9 | 24 |
| MAX | 30589 | 72.3 | 64137 | 17902 | 4281 | 1094 | 3.92 | 45.47 | 60.88 | 97.38 | 184.4 | 187 | 22.89 | 53.09 | 43.83 | 99.38 | 86.7 | 92.1 |
| MIN | 484 | 22.5 | 659 | 283 | 129 | 109 | 1.09 | 3.9522 | 34.96 | 95.85 | 131.9 | 36 | 3.14 | 0 | 2.68 | 62.25 | 51.7 | 2 |

**Figure 9**
Table 1.

### *Hi-fidelity sequence to structure alignment*

We modified the ClustalW software package [54] to initiate a global alignment of two neighbour sequences using the PAM series substitution matrices and apply position specific gap penalties by virtue of a secondary structure profile. The profile is derived from the structure's annotation information provided by the authors of the published structure as well as from NCBI's vector alignment search tool, VAST [55]. A greater weight is placed when the two sources agree, and this effectively forces gaps into unstructured regions lacking alpha helices and beta strands. To create conservative, fold-based alignments, gaps that were added to the genomic sequence are masked out since there is no correspondence to the structure and gaps inserted into the structure template to accommodate query insertions are eliminated (Figure 8). This gap-handling procedure had no visible effect on composition analyses that are later described.

To reject poor alignments and enhance the fidelity of the global alignment, both the sequence identity and structural position occupancy are determined over each VAST-identified structural domain. Various threshold levels were tested, although an alignment sequence identity of 25% and domain occupancy of 75% was found to provide optimal compromise between sensitivity and specificity (data not shown). If less than 25% of the aligned residues are identical and less than 75% of the aligned residues occupy residue positions in the domain, the domain is masked out completely and not used in any further computations (Figure 8). These selection criteria generate relevant domain homologues and provide the ability to discriminate subtle sequence changes that are independent of fold in a statistically observable manner. When an alignment across a domain is found to satisfy the minimum constraints specified above, a structural model is generated for the genomic sequence by virtue of a sequence-to-structure alignment, herein referred to as the model.

Since a genomic sequence may make many models using different templates, only the single best model is selected to minimize sampling bias. The selection criterion emphasizes the use of multi-domain structural models by using a scoring function derived from residue length of the non-masked out aligned domain region(s), the fraction of residues that are identical and the fraction of residues occupying domain positions. The model with the best score is then selected to represent that genomic sequence.

For each representative model, the sequence alignments between the genomic sequence and the template, along with the corresponding secondary structure are written to a database, herein known as the species-specific fold database. This fold database is the source of model and template sequences for determining fold composition and deriving species-specific fold scoring functions.

### *Model quality*

Our method exploits species-specific optimizations at the sequence level by making accurate structural-based alignment for genomic sequences. We generated models for 95 of the 100 genomes with 5 genomes having been very recent additions. Initially, there are as many 3D neighbours as genomic sequences (Table 1). However, 24 ± 10% of genomic sequences make structural models and only 19 ± 6% settle with a single representative model structure that pass our structural domain alignment criteria. The representative models are 168 ± 10 residues in length and possess 41 ± 5% sequence identity and 96.7 ± 0.3% domain occupancy with the template structure, which is clearly higher than our set minimum requirements. Furthermore, template structures are used 1.4 ± 0.4 times for model building, thereby minimizing structure over-sampling and providing more unique templates. Interestingly, at least 36 to as many as 287 different organisms contribute 5.8 ± 3.5 template structures to each genome modeling exercise. 30 ± 12% of the templates are obtained from *E. coli* and 23 ± 9% of structures are obtained from thermophilic species. Our models hold properties of 'good' models since they are based on at least 30% sequence identity are shorter than 200 residues and are aligned along template domains, in agreement with other published criteria [56].

In general, the number of final models generated for complete genomes reported elsewhere is greater than the number generated with our method. For example, the NCBI provides a substantially larger set of 3D structure neighbours for complete genomes, in which as many as 39% of sequences are reported to have structure neighbours [http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PDB_bact.html]. ModBase has on average between 2 to 4 models per sequence in which they claim roughly 44% are reliable [http://pipe.rockefeller.edu/modbase]. Since our comparative modeling method is more conservative in that it does not attempt to model side-chains, loops, or regions with no template and our alignments are evaluated over smaller, domain-focused regions, we expect fewer errors [57].

## Authors' contributions

KM provided the framework for complete genome analysis with her development of the SeqHound sequence and structure database management system. MD carried out the statistical analysis, derived and tested the scoring functions and drafted the manuscript. MD and CWVH jointly conceived of the study, and participated in its design and coordination.

## Tables

Table 1 – Summary statistics for complete genomes modeling and scoring function results. Each species is represented by a short abbreviation (Abbr.), a unique GenBank taxonomy identifier (TaxID), Class (A – Archae, B – Bacteria, C – Eukaryote), environmental optimization (Hyperthermophile HT, thermophile (T), halophile (H), acidophile (D), ureaphile (U), radiation resistant (R), intracellular pathogen (I), pathogen (P), solventogenic (S), symbionts (Y) and plant pathogen (PP)). The modeling statistics include the number of predicted open-reading frames (ORFs), the GC content of the predicted open-reading frames (%GC), number of sequence neighbours with 3D structures (N3D), the number of sequences with a potential to make a model (PM), the number of representative models selected (SM), the percentage of ORFs modeled (OM), the number of unique structure templates (UT), the number of times a template was used (TU), the average identity (%ID) and domain occupancy (%OCC) in sequence to structure alignments, and the average number of residues per model (Res). The taxonomic contribution is listed by the number of organisms that contributed template structures (OC), the average number of structures contributed by each (OF), the percentage of templates that were from E. coli (%E) and thermophiles (%T). Finally, the percentage of correctly identified sequences in jackknifing for the CF scoring function (CF(JK)) and the percentage of correctly identified sequences using the top 15 scoring scores for the CG scoring function (CG(P)) and for the CF scoring function (CF(P)). NA – not available.

## Additional material

### Additional File 1

*Species-specific genome composition (CG) scoring functions*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-39-S1.txt]

### Additional File 2

*Species-specific fold composition (CF) scoring functions*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-39-S2.txt]

## Acknowledgments

We would like to thank our colleagues at the Samuel Lunenfeld Research Institute for their support in our work. We would like to thank Gary Bader and Doron Betel for critical reading of the manuscript. We would also like to thank our reviewers for providing excellent feedback that enhanced the quality of our manuscript. This research was supported by grants to C.W.V Hogue and M. Dumontier by the Natural Sciences and Engineering Research Council of Canada.

## References

1. Martin DD, Ciulla RA, Roberts MF: **Osmoadaptation in archaea.** *Appl Environ Microbiol* 1999, **65:**1815-25
2. Gross M, Jaenicke R: **Proteins under pressure. The influence of high hydrostatic pressure on structure, function and assembly of proteins and protein complexes.** *Eur J Biochem* 1994, **221:**617-30
3. Vieille C, Zeikus GJ: **Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability.** *Microbiol Mol Biol Rev* 2001, **65:**1-43
4. Audia JP, Webb CC, Foster JW: **Breaking through the acid barrier: an orchestrated response to proton stress by enteric bacteria.** *Int J Med Microbiol* 2001, **291:**97-106
5. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V: **Complete genomic sequence of Pasteurella multocida, Pm70.** *Proc Natl Acad Sci U S A* 2001, **98:**3460-5
6. Oren A: **Bioenergetic aspects of halophilism.** *Microbiol Mol Biol Rev* 1999, **63:**334-48
7. Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D: **Molecular evolution of protein atomic composition.** *Science* 2001, **293:**297-300
8. Kreil DP, Ouzounis CA: **Identification of thermophilic species by the amino acid compositions deduced from their genomes.** *Nucleic Acids Res* 2001, **29:**1608-15
9. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, KD Linher, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, D Richardson, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, *et al*: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.** *Nature* 1999, **399:**323-9
10. She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, Erauso G, Fletcher C, Gordon PM, Heikamp-de Jong I, Jeffries AC, Kozera CJ, Medina N, Peng X, Thi-Ngoc HP, Redder P, Schenk ME, Theriault C, Tolstrup N, Charlebois RL, Doolittle WF, Duguet M, Gaasterland T, Garrett RA, Ragan MA, Sensen CW, Van der Oost J: **The complete genome of the crenarchaeon Sulfolobus solfataricus P2.** *Proc Natl Acad Sci U S A* 2001, **98:**7835-40
11. White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, Moffat KS, Qin H, Jiang L, Pamphile W, Crosby M, Shen M, Vamathevan JJ, Lam P, McDonald L, Utterback T, Zalewski C, Makarova KS, Aravind L, Daly MJ, Fraser CM, *et al*: **Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1.** *Science* 1999, **286:**1571-7
12. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ: **Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics.** *Microbiol Mol Biol Rev* 2001, **65:**44-79
13. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: **Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS.** *Nature* 2000, **407:**81-6
14. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S: **Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58.** *Science* 2001, **294:**2323-8
15. Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl TM, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter FJ, Weidner S, Wells DH, Wong K, Yeh KC, Batut J: **The composite genome of the legume symbiont Sinorhizobium meliloti.** *Science* 2001, **293:**668-72
16. Houry WA: **Mechanism of substrate recognition by the chaperonin GroEL.** *Biochem Cell Biol* 2001, **79:**569-77
17. Kim R, Kim KK, Yokota H, Kim SH: **Small heat shock protein of Methanococcus jannaschii, a hyperthermophile.** *Proc Natl Acad Sci U S A* 1998, **95:**9129-33

18. Mogk A, Tomoyasu T, Goloubinoff P, Rudiger S, Roder D, Langen H, Bukau B: **Identification of thermolabile Escherichia coli proteins: prevention and reversion of aggregation by DnaK and ClpB.** *Embo J* 1999, **18:**6934-49
19. Kowalski JM, Kelly RM, Konisky J, Clark DS, Wittrup KD: **Purification and functional characterization of a chaperone from Methanococcus jannaschii.** *Syst Appl Microbiol* 1998, **21:**173-8
20. Bock AK, Glasemacher J, Schmidt R, Schonheit P: **Purification and characterization of two extremely thermostable enzymes, phosphate acetyltransferase and acetate kinase, from the hyperthermophilic eubacterium Thermotoga maritima.** *J Bacteriol* 1999, **181:**1861-7
21. Russell RJ, Ferguson JM, Hough DW, Danson MJ, Taylor GL: **The crystal structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 A resolution.** *Biochemistry* 1997, **36:**9983-94
22. Lobry JR: **Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species.** *Gene* 1997, **205:**309-16
23. Lynn DJ, Singer GA, Hickey DA: **Synonymous codon usage is subject to selection in thermophilic bacteria.** *Nucleic Acids Res* 2002, **30:**4272-7
24. Chakravarty S, Varadarajan R: **Elucidation of determinants of protein stability through genome sequence analysis.** *FEBS Lett* 2000, **470:**65-9
25. Chakravarty S, Varadarajan R: **Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study.** *Biochemistry* 2002, **41:**8152-61
26. Tekaia F, Yeramian E, Dujon B: **Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis.** *Gene* 2002, **297:**51
27. Maes D, Zeelen JP, Thanki N, Beaucamp N, Alvarez M, Thi MH, Backmann J, Martial JA, Wyns L, Jaenicke R, Wierenga RK: **The crystal structure of triosephosphate isomerase (TIM) from Thermotoga maritima: a comparative thermostability structural analysis of ten different TIM structures.** *Proteins* 1999, **37:**441-53
28. Szilagyi A, Zavodszky P: **Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey.** *Structure Fold Des* 2000, **8:**493-504
29. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-71
30. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17:**849-50
31. Nielsen H, Engelbrecht J, S Brunak, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6
32. Lupas A, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252:**1162-4
33. Rost B, Fariselli P, Casadio R: **Topology prediction for helical transmembrane proteins at 86% accuracy.** *Protein Sci* 1996, **5:**1704-18
34. Chou KC, Maggiora GM: **Domain structural class prediction.** *Protein Eng* 1998, **11:**523-38
35. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105-32
36. Cai YD, Liu XJ, Xu XB, Chou KC: **Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect.** *J Cell Biochem* 2002, **84:**343-8
37. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF: **From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.** *Biotechnology (N Y)* 1996, **14:**61-5
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-402
39. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94
40. Uberbacher EC, Mural RJ: **Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach.** *Proc Natl Acad Sci U S A* 1991, **88:**11261-5
41. Gelfand MS: **Prediction of function in DNA sequence analysis.** *J Comput Biol* 1995, **2:**87-115
42. Dennis PP, Shimmin LC: **Evolutionary divergence and salinity-mediated selection in halophilic archaea.** *Microbiol Mol Biol Rev* 1997, **61:**90-104
43. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP: **Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.** *Nature* 2001, **414:**450-3
44. Clarke GD, Beiko RG, Ragan MA, Charlebois RL: **Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores.** *J Bacteriol* 2002, **184:**2072-80
45. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8
46. Ahern TJ, Klibanov AM: **The mechanisms of irreversible enzyme inactivation at 100C.** *Science* 1985, **228:**1280-4
47. Tomazic SJ, Klibanov AM: **Mechanisms of irreversible thermal inactivation of Bacillus alpha-amylases.** *J Biol Chem* 1988, **263:**3086-91
48. Fukuchi S, Nishikawa K: **Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria.** *J Mol Biol* 2001, **309:**835-43
49. Bryant SH, Lawrence CE: **The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions.** *Proteins* 1991, **9:**108-19
50. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of the National Center for Biotechnology Information: 2002 update.** *Nucleic Acids Res* 2002, **30:**13-6
51. Michalickova K, Bader GD, Dumontier M, Lieu HC, Betel D, Isserlin R, Hogue CW: **SeqHound: biological sequence and structure database as a platform for bioinformatics research.** *BMC Bioinformatics* 2002, **3:**32
52. Dumontier M, Hogue CWV: **NBLAST: a Cluster Variant of BLAST for NxN Comparisons.** *BMC Bioinformatics* 2002, **3:**13
53. Wang Y, Bryant S, Tatusov R, Tatusova T: **Links from genome proteins to known 3-D structures.** *Genome Res* 2000, **10:**1643-7
54. Higgins DG, Sharp PM: **CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.** *Gene* 1988, **73:**237-44
55. Hogue CW, Ohkawa H, Bryant SH: **A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database.** *Trends Biochem Sci* 1996, **21:**226-9
56. Melo F, Sanchez R, Sali A: **Statistical potentials for fold assessment.** *Protein Sci* 2002, **11:**430-48
57. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes.** *Annu Rev Biophys Biomol Struct* 2000, **29:**291-325