A General Horizontal Alignment Tool for Shape Comparison of Diverse One-Dimensional Protein Data

Omar Hadzipasic, James O. Wrabl, and Vincent J. Hilser

Abstract

A general algorithm is presented that returns the optimal pairwise gapped alignment of two sets of signed numerical values. One distinguishing feature of this algorithm is a flexible comparison engine (based on both relative shape and absolute similarity measures) that does not rely on explicit gap penalties. Importantly, an empirical probability model is developed to estimate the significance of the returned alignment with respect to a particular type of random data. The utility of the algorithm for database search and pairwise alignment is demonstrated on diverse types of protein and nucleic acid data, including average side-chain hydrophobicity, native-state thermodynamic stability, and mRNA codon translation efficiency. Results of biological interest are obtained in each case. One final example, possible remote homology between the *Chlamydia TC0624 Inc* protein and the pore-forming domain of colicin indicates that the algorithm can inform medical discovery, complementing existing protein sequence and structure based tools. The source code, documentation, and a basic web-server application are freely available at http://best.bio.jhu.edu/HePCaT/.

Introduction

Many penetrating insights into protein function and evolution have been inferred from analysis of amino acid sequences (1, 3, 5, 46) or comparison of three dimensional atomic structures.(20, 40, 41, 44, 45) However, protein function and evolution arise from a manifold of physical, chemical, and biological mechanisms, only partly accounted for by side chain identity or structure similarity.(9, 16, 27, 32, 49) Consequently, proteins can and should be meaningfully characterized by other attributes, such as the energetic contributions to stability (15) or the predicted codon translation efficiency along the mRNA. (14, 51) Such attributes are not easily accommodated by simple adaptation of current algorithms, largely because the scoring systems for such algorithms are based on positional sequence identity (amino acid substitution matrices) or absolute geometric structural similarity (Euclidean distance).

The resulting unfortunate situation is that properties other than sequence and structure, and their additional potential biological insight into proteins, have not been as thoroughly explored.

For example, the local thermodynamic stability of a protein, as experimentally measured by deuterium-hydrogen exchange (7, 28), is described by a one-dimensional sequence of numerical values (*i.e.* amide protection factors). These values are well-known to be a combination of sequence, structure, and solvent effects (6), but no substitution matrix or distance measure exists for the objective comparison of two sets of protection factors. Important knowledge might be missed due to the inability to make such comparisons. Worse, erroneous conclusions might be inferred from comparisons that separate the effects (*e.g.* comparing side chain identity in the absence of information about the thermodynamic stability at the same position).

One-dimensional software tools have been developed for the special case of hydrophobicity analysis, such as identification and alignment of the membrane spanning regions of non-globular proteins. (11, 24, 30) While useful, these tools have historically incorporated family-specific scoring matrices (17) and empirical gap penalties. Such heuristics hinder the algorithms' transferability to different proteins or applicability to data types other than transmembrane protein hydrophobicity. In addition, the scoring functions for hydrophobicity analysis are often based on absolute similarity (29), and while this is effective at finding matches that are similar in both shape and magnitude, two sets of data that describe the same shape, but are offset by a constant value, would be missed. For example, such a situation can arise for experimentally measured local thermodynamic stabilities of proteins, where the relative stabilities of the same structural region of two homologs are observed to be strikingly similar, yet offset by a constant ΔΔG value (18). Finally, some of these previous tools lack the capability for large database searches or do not include estimates of statistical significance, limiting their usefulness even for the appropriate input data.

To address these shortcomings, we have developed a general tool to compare the one-dimensional profiles defined by arbitrary sequences of numerical data. To maximize the flexibility of the tool, we have deliberately chosen in the design to include two metrics that match both the relative shapes of the two profiles as well as the absolute similarity of the numerical values. Thus, the scoring system is designed to be independent of the input data type, and its utility is demonstrated on three diverse types of protein data normally not analyzable with a single software package. Because such a design emphasizes the closeness in shape of the two sets scanned over a horizontal range of positions, in contrast to the vertical position-by-position independent scoring of a standard amino acid substitution matrix, the algorithm is named *Horizontal Protein Comparison Tool* (*HePCaT*).

Materials and Methods

Detailed description of the HePCaT algorithm

The algorithm proceeds by creating internal signed distance matrices from each of two sets of input numerical data vectors (Figure 1, Step 1). For a protein of *M* residues, each element of its distance matrix ***D*** is defined as

$$D_{i=1..M,j=1..M} = sign(v_i - v_j)\sqrt{(v_i - v_j)^2} \quad (1)$$

The signed distance matrices, while not symmetric, are mirror images across the diagonal (Figure 1, Step 2). Thus, both shape and magnitude information about each data set are encoded in these matrices. For example, the Protein 2 matrix $\boldsymbol{D_2}$ (Figure 1, Step 2) clearly indicates the strong local maximum in the N-terminal half relative to the strong local minimum in the C-terminal half as prominent red or blue regions.

Equation 1 demonstrates a key conceptual difference from structure comparison algorithms that are usually based on distance or contact matrices restricted to only positive values (19, 43). This difference reflects the nature of the information being compared. For structure comparison, the distance between two atoms is identical whether it is computed between the first and second atom or *vice versa*, while in the case of thermodynamic stability, for example, there may be a relative stabilization between the first and second atoms which becomes a relative destabilization between second and first. The sign in Equation 1 thus represents this key conceptual difference: a distance in *HePCaT* has *both* sign and magnitude. (It is noted that Equation 1 may be generalized to an arbitrary number of dimensions, but the present work only considers the one-dimensional case.)

A shape similarity matrix, ***S***, is then constructed from the two distance matrices (Figure 1, Step 3). To speed the calculation, a heuristic window size, *W*, is introduced. (In the present work, *W* is always five residues, but we note that this is potentially an adjustable parameter and a completely exhaustive search may be performed with *W* = 1.) For each position $i = M - W - 1$ in Protein 1 and each position $j = N - W - 1$ in Protein 2, the relative shape similarity is computed between the two five-residue blocks originating at positions *i* and *j* :

$$S_{i=1..M-W-1,j=1..N-W-1} = \frac{1}{W}\sum_{k=0}^{W-1}\left|D_{1i,i+k} - D_{2j,j+k}\right| \quad (2)$$

Equation 2 is simply the average absolute value of the difference of equivalenced internal distances between the two blocks. If the shape similarity is high this value will be small, if the shape similarity is very different this value will be large. Such dissimilarity can be readily viewed for the example proteins: the Figure 1 similarity matrix contains strong positive values (darkest

red) where the large peak in the middle of the first protein coincides with the deep valley in the C-terminal region of the second (or *vice versa*).

In this work, the signed internal distances within each block of W = 5 residues are scaled such that the longest absolute value of the internal distance is one,

$$D_{i,j+k} = \frac{D_{i,i+k}}{\max(abs(D_{i,i+k})\,|_{k=1}^{k=W})} \qquad (3)$$

Although this normalization can be disabled, we believe that emphasizing comparison of relative shape improves detection of trends in biological data, which can exhibit wide variations in scale. Normalization also facilitates the choice of the user-defined alignment shape similarity cutoff, as described below.

The optimal alignment between Proteins 1 and 2 is found by exhaustive search of the shape similarity matrix (Figure 1, Steps 4 and 5). "Optimal" is defined as the largest unique set of blocks of size W, subject to at most *GapMax* skipped positions of the similarity matrix between blocks, which exhibits the smallest *RMSD* of all such sets passing a user-defined shape similarity cutoff, C. If C = 0, exact shape matches only are permitted in the alignment list. For this work, where Equation 3 applies, C was set to 0.40, meaning that an alignment whose average normalized distance between two five residue blocks was at most 40% different was counted as a matching shape. If Equation 3 was disabled, C would have to be adjusted empirically based on the dynamic ranges of data compared.

The algorithm starts at cell (1,1) of $\boldsymbol{S}$ *(i.e.* the lower left corner of the matrix in Figure 1, Step 3), corresponding to the average difference between the scaled intraprotein distances of residues 1 – 5 in Protein 1 and residues 1 – 5 in Protein 2. If $\boldsymbol{S}_{1,1} <= C$, this match is kept and position $\boldsymbol{S}_{6,6}$ is checked, until all cells of $\boldsymbol{S}$ are evaluated up to the position $\boldsymbol{S}_{M-W+1,N-W+1}$ (*i.e.* the upper right corner of the matrix in Figure 1, Step 3). If at any point $\boldsymbol{S}_{i,j} > C$, single cell gaps are inserted in one or both sequences up to a maximum of *GapMax* in an attempt to obtain the longest path through $\boldsymbol{S}$ subject to C. A list of the longest gapped paths is kept at this stage (Figure 1, Step 3, colored arrows). Therefore, all paths in this list are comprised of equivalenced positions in the two proteins such that, on average, the intraprotein distances seen at every position match to at least degree C; this average value is named Average Path Distance (*APD*, Figure 1, Step 4). *GapMax* was empirically set to 4 for this work. No penalty is applied to *APD* for insertion of a gap. Importantly, at this first stage only relative shape similarity is checked; any systematic offset between the two data sets is ignored because only the differences between intraprotein distances are evaluated.

After ***S*** has been exhaustively searched, the list of longest alignments passing the shape cutoff is filtered by *RMSD* of the aligned residues (Figure 1, Step 5). The smallest *RMSD* alignment is defined as the optimal (thus, the *RMSD* is a magnitude filter). If multiple alignments of identical longest length happen to exhibit identical *RMSD*, only the first such one encountered is returned. In *HePCaT*, the *RMSD* calculation is executed after translation of both sets to data to their respective centers-of-mass, thus effects of a global offset between each data set are again minimized. Following Jia, *et al.* (22), we assign an Optimal Path Score (*OPS*) to this optimal alignment according to the formula:

$$OPS = \frac{RMSD}{L}\left(1+\frac{Gaps}{L}\right) \qquad (4)$$

In Equation 4, *L* is the alignment length and *Gaps* is the total number of cells skipped in ***S*** to obtain that alignment. Note that gaps are not explicitly penalized during alignment, but gaps will penalize the final score according to Equation 4, under the reasonable and common assumption that a gapless match is a "better" match than a gapped one. Of course, the *GapMax* parameter can be set to zero if desired so that all gaps are forbidden.

Probability models to estimate the significance of an *OPS* score *s* of an alignment of length *L* were derived from analysis of randomly generated alignments (Figure 1, Step 6). It is important to note that these probability models are specific to the type of protein data aligned and must be recalibrated for a specific combination of *W*, *C*, and *GapMax*. More details about the probability models for the three types of data analyzed in this work are given below. Probability models for these data (Kyte-Doolittle hydropathy (25) averaged over a 9-residue window, eScape predicted native state thermodynamic stability (15), and predicted translation efficiency index tAI (14, 51) averaged over a 9-residue window) were built for the following *HePCaT* parameters as listed in Supplementary Material: *W* = 5 residues, *GapMax* = 4 residues, and *C* = 0.4 with the local scaling given in Equation 3.

Construction of probability models

Significance of the Equation 4 score of optimal *HePCaT* alignments was estimated with respect to random optimal alignments of identical length. Two random proteins of equal lengths between 10 and 500 residues were generated according to background amino acid frequencies as given by Robinson & Robinson. (39) Sets of at least 20,000 such pairs were optimally aligned using *HePCaT*, and the distributions of Equation 4 scores for a given optimal alignment length were tabulated (Figure 2A). It was observed that these skewed unimodal distributions exhibited a strong dependence on alignment length. Out of several possible two-variable

formulae, it was empirically determined that these score distributions were statistically best fit by scaled inverse chi-square probability density functions (Figure 2A, Supplementary Table S1) (23),

$$PDF_{InverseChiSquared}\left(x;\nu,\sigma^2 \mid L\right)=\frac{\left(\frac{\sigma^2\nu}{2}\right)^{\frac{\nu}{2}} e^{-\frac{\sigma^2\nu}{2x}}}{\Gamma\left(\frac{\nu}{2}\right)x^{1+\frac{\nu}{2}}} \tag{5}$$

In Equation 5, *L* is optimal alignment length, and Γ(x) is the Gamma function. (36) Parameters ν and σ^2 were estimated by minimum chi-squared fits to the binned score data at each observed alignment length (Figure 2A). Binning and parameter estimation were performed using custom *Mathematica* 8.0 scripts, such that each variable-width bin contained at least 20 points, additional details are provided in Supplementary Table S1.

Ad-hoc analytical expressions were fitted to the collected best-fit parameters of Equation 5 as a function of optimal alignment length *L* (Figure 2B):

$$\nu\left(L \mid W,C,GapMax\right)=m\left(L\right) \tag{6}$$

$$\sigma^2\left(L \mid W,C,GapMax\right)=e^{a+b\ln(L+c)} \tag{7}$$

Determination of coefficients *a*, *b*, *c*, and *m* only employed reasonably well-fit Equation 5 values whose null hypotheses (*i.e.* that the simulated data were drawn from Inverse Chi Square Distributions) could not be rejected at $p < 0.05$. Equations 6 and 7 coefficients for the various biological data sets used in this work are given in Table 1, all resulted from excellent fits of R^2 = 0.99 or better using *gnumeric* spreadsheet software (Figure 2B).

Therefore, given an observed optimal *HePCaT* alignment of length *L* with Equation 4 score *s*, the probability *p* of observing that alignment by chance could be estimated from the corresponding scaled inverse chi-square cumulative distribution function as:

$$p\left(s \mid L,W,C,GapMax\right)=\int_0^{x<s} CDF_{InverseChiSquared}\left(x;\nu(L),\sigma^2(L)\right)=\int_0^{x<s} Q\left(\frac{\nu}{2},\frac{\sigma^2\nu}{2x}\right) \tag{8}$$

In Equation 8, *Q(a,x)* is the complement of the regularized Gamma function (36); ν and σ^2 were estimated from Equations 6 and 7, using coefficients specific to the particular biological data set under consideration.

Hydropathy database search of the human proteome using adenosine receptor A2a as query

The human proteome was obtained from translation of the DNA sequences contained in the NCBI CDDS (37) build 36.3 (April 30, 2008). Each amino acid in every protein was assigned a side-chain hydrophobicity value according to the Kyte-Doolittle hydropathy scale. The values for each protein were averaged using a nine-residue sliding window; averaged values for the first and last four residues in each protein were subsequently ignored. The averaged values for human adenosine receptor A2a (CCDS 13826.1, gi|5921992) were used as query to the human proteome, *i.e.* the averaged hydropathy values of each protein in the proteome were optimally pairwise aligned to A2a using *HePCaT* with the following parameters: *W* = 5 residues, *C* = 0.4, *GapMax* = 4 residues. *P*-values for each alignment were computed using a probability model specific to these data (Table S1) and Table 1 parameters as described above. GPCRs were annotated in the human proteome by *FASTA*-aligning (34) amino acid sequences of the proteome with amino acid sequences of known GPCRs obtained from the GPCRDB. (52)

Pairwise alignment of disordered N-terminal glucocorticoid receptor domains based on predicted stability

A BLAST (2) search of the NCBI *nr* database (12/19/11, 16,645,108 sequences) with the full-length human glucocorticoid receptor protein (GR, gi|121069, 777 letters) as query was performed on the NCBI website using default parameters. 99 significant hits were returned. Clustering of the hits at 90% identity using cd-hit (21) with otherwise default parameters and removal of one partially redundant GFP-chimera sequence resulted in 24 unique proteins. A multiple sequence alignment of these 24 proteins was computed using *PROMALS3D* (35) with default parameters (Supplementary Material, Figure S1). The amino acid sequences of the N-terminal domains (NTD) of each protein indicated by this multiple alignment were separately extracted. Each N-terminal subsequence was input to the *eScape* software (15), a package that predicts native-state local thermodynamic stability (ΔG) of a protein under physiological conditions, based on amino acid sequence. The *eScape* stability profiles of each NTD were then each pairwise realigned to the human GR NTD using *HePCaT* with the following parameters: *W* = 5 residues, *C* = 0.4, *GapMax* = 4 residues. *P*-values for each alignment were computed using a probability model specific to these data (Table S1) and Table 1 parameters as described above.

Pairwise alignment of homologous E. coli mRNA by predicted translation efficiency tAI

Putatively homologous proteins of *E. coli* were extracted from the *SCOP* database v1.73 (4, 33) by first matching all annotations of organism ("Escherichia coli") and then grouping non-

redundant members by identical class, fold, superfamily, and family. To accurately map the *SCOP* amino acid sequence to the *CSANDS* (42) mRNA, identical amino acid sequences, as aligned by *FASTA* between this initial set and the *CSANDS* database, were also manually inspected to ensure non-redundancy within families and retained for further analysis. The mRNA for each *E. coli* protein retained was obtained from *CSANDS*, and each mRNA codon of each sequence was assigned an estimated translation efficiency value, *tAI*, according to the values for *E. coli* given in Tuller, *et al*. (51) The *tAI* values for each mRNA were averaged using a nine-codon sliding window, averaged values for the first and last four codons in each protein were subsequently ignored. A total of 337 *E. coli* mRNAs from 128 *SCOP* families were ultimately analyzed. All pairs of mRNAs from putatively homologous proteins, 377 nonredundant pairs total, were pairwise aligned using *HePCaT* parameters of W = 5 codons, C = 0.4, *GapMax* = 4 codons. *P*-values for each alignment were computed using a probability model specific to these data (Table S1) and Table 1 parameters as described above. For comparison, Kyte-Doolittle hydropathy profiles were also constructed for these 337 proteins, the 377 pairs were pairwise aligned using *HePCaT*, and *p*-values specific to hydropathy were computed as described above for the human proteome.

Discovery of similarity between ORFan protein TC0624 and colicin pore-forming domain

A dataset of 8812 ORFan protein sequences was obtained from Yomtovian, *et al.(54)* As described above, *HePCaT* was used to optimally align the Kyte-Doolittle averaged hydropathy profiles of each ORFan protein with the profile of each member of a non-redundant set of 214 membrane proteins of known structure derived from the ASTRAL domain database.(10) These 214 proteins were the representatives resulting from a 70% sequence identity cd-hit (21) clustering of all membrane proteins (class *f*) in the SCOP 1.73 database.(4) Secondary structure prediction was performed using the Jpred3 server (12) and Hidden Markov Model sequence profile comparison was performed using the HHpred server(47), both with default parameters.

Results

The utility of *HePCaT* was assessed by exploring three different biological questions: hydrophobicity similarity search against a database, pairwise alignment of local thermodynamic stability, and conservation of translation efficiency in *E coli*. Results described below provided biological insight from these common bioinformatics tasks, while simultaneously illuminating the strengths and weaknesses of the algorithm's design.

Database search using human adenosine receptor A2a as query

The hydropathy profile of the human adenosine A2a 7Tm G-protein coupled receptor (GPCR) was used to search the human proteome for close matches based solely on hydrophobicity patterns. As expected, hundreds of known 7Tm GPCRs were significantly matched by *HePCaT* ($p < 0.01$, data not shown). The most significant ten matches are displayed in Figure 3. These hits clearly fell into two categories: those that matched the transmembrane region (50) of A2a (Figure 3, blue) and those that matched the tail region (Figure 3, red). The longest match to the transmembrane region was the A2b isoform, which is also 59% sequence identical to A2a. Unexpectedly, a Type 2 taste receptor also exhibited a significant match to this region (Figure 3). As this taste receptor has undetectable pairwise sequence identity to A2a and its structure has not been experimentally determined, this observed similarity may be a useful template for a homology model based on the A2a structure. (48)

We attempted to rationalize the best matches to the A2a tail region in terms of sequence, structure, or function. However, in contrast to the transmembrane region matches, biological explanations for these remain mysterious. Some of the proteins in this group are medically important, such as the hematological and neurological expressed-1 like protein, ephrin A4 isoforms, and the B and T-lymphocyte attenuator precursor. Structural information about some of these hits could not be confidently transferred to the putatively disordered tail region of A2a, which is thought to be involved in ligand specificity of the GPCR. (26) The shortest hit to the tail region was possibly a statistical artifact: this metallothionein is naturally short and contains a high frequency of cysteine residues; such low-complexity sequences are normally filtered out of amino acid sequence searches (53), which was not done in the present study.

Pairwise alignment of disordered N-terminal domains of glucocorticoid receptor

A second pilot study using *HepCaT* involved pairwise alignment of the disordered N-terminal domains (NTD) of protein sequences homologous to the human glucocorticoid receptor (GR). These nonredundant sequences, found by objective *BLAST* search, exhibited significant sequence identity over their entire lengths and came from mammals, amphibians, and fish. A state-of-the-art multiple alignment of the full-length sequences clearly demonstrated the weaker sequence similarity in the N-terminal regions relative to C-terminal regions (Figure S1), and the consequently lower confidence in the positional correctness of the N-terminal alignment. As the NTDs are known to mediate ligand specificity and biological activity of GR, we wished to use additional information about the estimated thermodynamic stability to possibly reveal important functional insights not obtainable from the less reliable sequence comparisons. The locally stable and unstable regions of each NTD could be represented as

"peaks" and "valleys" and, like average hydrophobicity, were thus amenable to optimal pairwise comparison using *HePCaT*.

Each NTD stability dataset was separately aligned to the human NTD and significance of each comparison was computed as described in Methods. These alignments are displayed in Figure 4, arranged such that the most significant *HePCaT* comparisons are closer (top) to the human query and the least significant comparisons furthest away (bottom). Two important results were observed: first, NTDs from warm-blooded organisms (red) exhibited more significant thermodynamic similarity to the human NTD, and second, aligned regions of thermodynamic similarity generally often corresponded to one or more functionally relevant regions of the NTD, the so-called "AF1" and "scaffold" regions. These results suggest that NTD thermodynamic properties of mammals are significantly different than those of fish and amphibians. The functional interactions between isolated and intact human GR domains are currently under active study, and these predicted thermodynamic differences in the NTD may have biological implications. In particular, priority could be given to investigation of the isolated AF1 region of *A. carolinensis* and the scaffold region of *X. tropicalis*, as they seem to be the non-mammailian homologs with the most similar local stabilities to human.

Conservation of predicted mRNA translation efficiency of homologous E. coli proteins

A third application of the *HePCaT* algorithm was to answer the question: is the positional (codon-specific) translation speed of an mRNA conserved? To address this issue, more than 300 pairs of proteins highly similar in sequence and structure were extracted from the *E. coli* proteome according to the expert-curated classifications in the *SCOP* database (Methods). Crucially, proteins belonging to the same *SCOP* family are likely to be homologous, that is, descended from a common ancestor and thus, likely to be evolutionarily conserved. mRNA coding for each homologous protein was obtained from the *CSANDS* database and the predicted translation efficiency at each codon was computed according to the *tAI* values of Tuller, *et al.* These *tAI* values are thought to be a reasonable measure of translation speed through the ribosome at the codon level. The locally faster and slower regions along each mRNA could be represented as "peaks" and "valleys" and were thus amenable to optimal pairwise comparison using *HePCaT*.

Each homologous protein pair's *tAI* values were aligned and significances computed. When the p-values for each alignment were tabulated, a surprising result emerged: *p*-values for translation efficiency were rather evenly distributed across all possible values between zero and one (Figure 5, dark curve). This implied that, for most pairs of homologous mRNA, any similarities in translation efficiency were not significantly different from randomly uniform

distribution. As a control, similarities in hydrophobicity for the same protein pairs showed a skewed distribution, with approximately one-third of all pairs exhibiting moderately significant similarity at $p < 0.10$ (Figure 5, light curve). Thus, it was concluded that position-specific translation efficiency, in contrast to hydrophobicity, sequence, or structure similarity, is not an evolutionarily conserved property of proteins.

Predicted remote homology between the pore forming domain of colicin and Chlamydia TC0624 protein

One final example of the utility of *HePCaT* concerns the possible discovery of remote homology with medical importance. The *C. muridarum* protein *TC0624*, classified as an "ORFan" due to the absence of sequence similarity between any other known proteins,(54) nonetheless exhibited a significant *HePCaT* hydropathy match to the pore forming domain of *E. coli* colicin A (Figure 6A). Secondary structure prediction was consistent with tertiary structural similarity (Figure 6A), and sensitive sequence search using Hidden Markov Models revealed marginal, but repeated, similarity to the sequence of colicin implicated in the hydropathy match (Figure 6B). Thus, a total of four lines of evidence (hydropathy, secondary structure prediction, sequence similarity, and the regional correspondence between the sequence and structure matches) all converged on similarity between *TC0624* and the pore forming domain of colicin. However, this conclusion would have not been possible without the original significance of the *HePCaT* hydropathy match.

Importantly, the hydrophobic region of colicin implicated in this match has long been thought to be functionally crucial for colicin's lethal ability to travel from a hydrophilic extracellular environment, insert into the hydrophobic membrane interior, and form toxic pores in its host.(13) *TC0624* has independently been placed (31) in a class unique to *Chlamydiae* that is observed by experiment to also similarly partition into the membrane interior of the chlamydial inclusion.(8) These so-called "*Inc*" proteins, difficult or impossible to predict using existing computational tools,(31) are nonetheless important for chlamydial survival and maturation within its human and animal hosts. It appears that the extreme hydrophobicity exhibited by the *inc* proteins (8) permits their computational prediction using *HePCaT*. A novel functional hypothesis for these medically important proteins is also suggested: the *Incs* may form membrane-spanning pores that obtain nutrition from the host cytoplasm. Finally, this example demonstrates that this "ORFan" may actually belong to a known protein family.

Discussion

Most protein and nucleic acid data contained within the avalanche of next-generation genome sequencing can be expressed as sequentially numeric "peaks" and "valleys". These data include, but are not limited to, gene expression, ribosomal profiling, *ChIP Seq*, *RNASeq*, mRNA translation efficiency, thermodynamic stability of protein or mRNA, and physico-chemical properties such as hydrophobicity. A gap exists among software algorithms for analysis of such data, and the *HePCaT* algorithm described in this work is designed to help fill this gap. To facilitate such analysis and discovery, a webtool that allows execution of the algorithm, visualization of the result, access to the raw and analyzed data, and download of the source code is available at http://best.bio.jhu.edu/HePCaT.

There are at least three distinguishing features of the *HePCaT* algorithm. First, the input is completely arbitrary: if the data can be expressed in numeric form regardless of its source, patterns can potentially be detected. Second, its generalized scoring system is sensitive to both shape and magnitude similarity, allowing some degree of pairwise alignment flexibility. Third, the *W* parameter emphasizes a horizontal matching of patterns, as contrasted with the vertical matching that commonly occurs with amino acid substitution matrices or profile PSSMs.

In our view, vertical evolutionary conservation of amino acids has been thoroughly explored using tools such as *BLAST* and *FASTA*, while horizontal conservation of other protein properties has not. Thus, non-local properties of proteins, depending on correlations across residue positions, such as thermodynamic stability, can now be potentially explored. Indeed, the Figure 4 findings of thermodynamic similarity between mammalian GR homologs could *never* have been detected using traditional amino acid sequence or structure-based analysis.

Rigorous evaluation of the statistical significance of a result is an essential piece of scientific data that is often neglected in bioinformatics tools. Indeed, the conclusions of the Figure 5 results could not have been obtained without the associated statistical significance. As with other tools, the *HePCaT* statistical significances require calibration specific to the input data and algorithm parameters. Although recalibration for random simulation data not covered by Table 1 parameters is straightforward, an alternative estimate of statistical significance is available. Specifically, the non-parametric statistics of the *MIC* score reported by Reshef, *et al.* (38) can be used to evaluate a match returned by *HePCaT*. We believe that the applicability of the *MIC* statistics is maximized with *HePCaT* parameters of *GapMax* = 0 and *W* = 1.

Acknowledgements

Grant support from the NSF (MCB-0446050) and NIH (GM063747) is gratefully acknowledged.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403-410
2. Altschul SF, Madden TF, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of database search programs. *Nucleic Acids Res* 25: 3389-3402
3. Alva V, Remmert M, Biegert A, Lupas AN, Soding J. 2010. A galaxy of folds. *Protein Sci* 19: 124-130
4. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36: D419-425
5. Aravind L, Koonin EV. 1999. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J Mol Biol* 287: 1023-1040
6. Bai Y, Milne JS, Mayne L, Englander SW. 1993. Primary structure effects on peptide group hydrogen exchange. *Proteins* 17: 75-86
7. Bai Y, Milne JS, Mayne L, Englander SW. 1994. Protein stability parameters measured by hydrogen exchange. *Proteins* 20: 4-14
8. Bannantine JP, Griffiths RS, Viratyosin W, Brown WJ, D.D. R. 2000. A secondary structure motif predictive of protein localization to the chlamydial inclusion membrane. . *Cellular Microbiology* 2: 35-47
9. Bryan PN, Orban J. 2010. Proteins that switch folds. *Curr Opin Struct Biol* 20: 482-488
10. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, et al. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res* 32: D189-D192
11. Clements JD, Martin RE. 2002. Identification of novel membrane proteins by searching for patterns in hydropathy profiles. *European Journal of Biochemistry* 269: 2101-2107
12. Cole C, Barber JD, Barton GJ. 2008. The Jpred3 secondary structure prediction server. *Nucleic Acids Research* 36: W197-W201
13. Cramer WA, Heymann JB, Schendel SL, Deriy BN, Cohen FS, et al. 1995. Structure-function of the channel-forming colicins. *Annual Reviews of Biophysics and Biomolecular Structure* 24: 611-641
14. dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* 32: 5036-5044
15. Gu J, Hilser VJ. 2008. Predicting the energetics of conformational fluctuations in proteins from sequence: a strategy for profiling the proteome. *Structure* 16: 1627-1637
16. Henzler-Wildman K, Kern D. 2007. Dynamic personalities of proteins. *Nature* 450: 964-972
17. Hill JR, Kelm S, Shi J, Deane CM. 2011. Environment specific substitution tables improve membrane protein alignment. *Bioinformatics* 27: 15-23
18. Hollien J, Marqusee S. 1999. Structural distribution of thermodynamic stability in a thermophilic enzyme. *Proceedings of the National Academy of Sciences, USA* 96: 13674-13678
19. Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* 233: 123-138
20. Holm L, Sander C. 1997. Dali/FSSP classification of protein folds. *Nucleic Acids Res* 25: 231-234
21. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26: 680-682
22. Jia Y, Dewey TG, Shindyalov IN, Bourne PE. 2004. A new scoring function and associated statistical significance for structure alignment by CE. *J Comp Biol* 11: 787-799
23. Johnson NL, Kotz S, Balakrishnan N. 1994. *Continuous Univariate Distributions*. New York, New York: John Wiley & Sons

24. Khafizov K, Staritzbichler R, Stamm M, Forrest LR. 2010. A study of the evolution of inverted-topology repeats from Leu T-fold transporters using AlignMe. *Biochemistry* 49: 10702-10713
25. Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157: 105-132.
26. Lebon G, Warne T, Edwards PC, Bennett K, Langmead CJ, et al. 2011. Agonist-bound adenosine A(2A) receptor structures reveal common features of GPCR activation. *Nature* 474: 521-525
27. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, et al. 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science* 001: 001-001
28. Liu T, Pantazatos D, Li S, Hamuro Y, Hilser VJ, Woods VLJ. 2012. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. *Journal of the American Society for Mass Spectrometry* 23: 43-56.
29. Lolkema JS, Slotboom DJ. 1998. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Molecular Membrane Biology* 15: 33-42
30. Lolkema JS, Slotboom DJ. 1998. Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins. *FEMS Microbiology Reviews* 22: 305-322
31. Lutter EI, Martens C, Hackstadt T. 2012. Evolution and conservation of predicted inclusion membrane proteins in chlamydiae. *Comparative and Functional Genomics* 362104: 1-13
32. Murzin AG. 2008. Metamorphic Proteins. *Science* 320: 1725-1726
33. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247: 536-540
34. Pearson WR. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 132: 185-219
35. Pei J, Kim BH, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res* 36: 2295-2300
36. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical recipes in C: the art of scientific computing*. New York: Cambridge University Press
37. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, et al. 2009. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research* 19: 1316-1323
38. Reshef DN, Reshef YA, Fuinucane HK, Grossman SR, McVean G, et al. 2011. Detecting novel associations in large data sets. *Science* 334: 1518-1524
39. Robinson AB, Robinson LR. 1991. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proceedings of the National Academy of Sciences, USA* 88: 8880-8884.
40. Sadreyev RI, Kim BH, Grishin NV. 2009. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol* 19: 321-328
41. Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9: 56-68
42. Saunders R, Deane CM. 2010. Synoymous codon usage influences the local protein structure observed. *Nucleic Acids Res* 38: 6719-6728
43. Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11: 739-747
44. Shindyalov IN, Bourne PE. 2000. An alternative view of protein fold space. *Proteins* 38: 247-260
45. Skolnick J, Arakaki AK, Lee SY, Brylinski M. 2009. The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci U S A* 106: 15690-15695
46. Soeding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21: 951-960

47. Soeding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33: W244-W248
48. Tebben AJ, Schnur DM. 2011. Beyond rhodopsin: G protein-coupled receptor structure and modeling incorporating the beta2-adrenergic and adenosine A2A crystal structures. *Chmoinformatics and computational chemical biology: methods in molecular biology* 672: 359-386
49. Tokuriki N, Tawfik DS. 2009. Protein dynamism and evolvability. *Science* 324: 203-207
50. Topiol S, Sabio M. 2009. X-ray structure breakthroughs in the GPCR transmembrane region. *Biochemical Pharmacology* 78: 11-20
51. Tuller T, Carmi A, Vestsigian KN, S., Dorfan Y, Zaborske J, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 14: 344-354
52. Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, et al. 2010. GPRCDB: information system for G-protein coupled receptors. *Nucleic Acids Res* 39: D309-D319
53. Wootten JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computational Chemistry* 17: 149-163
54. Yomtovian I, Teerakulkittipong N, Lee B, Moult J, Unger R. 2010. Composiiton bias and the origin of ORFan genes. *Bioinformatics* 26: 996-999

Tables

Table 1. Parameters used in Equations 6 and 7 to estimate random protein data probability distributions based on the Inverse Chi-Squared Distribution.

Data Type	***m***	***a***	***b***	***c***
Hydrophobicity	0.497609	0.160379	-1.04167	38.9045
Thermodynamic Stability	0.740239	9.11826	-1.23341	45.3881
E. coli tAI Value	0.455869	-2.96918	-0.961557	27.6082

Figure Legends

Figure 1. Overview of the Horizontal Protein Comparison Tool (*HePCaT*) algorithm. The hydrophobicity profiles of two proteins, each of length $M = N = 20$ residues, are shown (Step 1). Intraprotein signed distances are computed within each protein according to Equation 1 in the main text (Step 2). Positive distances, *e.g*. measured from a residue with a local minimum value to a residue with a local maximum value, are indicated in red, negative distances in blue. The signed distance matrices are therefore square and symmetrically reflected across the diagonal. Distances for protein 1 and protein 2 correspond to matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, respectively. The similarity matrix $\boldsymbol{S}$ that ultimately compares the two proteins is constructed from the average absolute distance differences of $W = 5$ residue blocks between $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, according to Equation 2 (Step 3). In $\boldsymbol{S}$, light colored squares indicate blocks of $W = 5$ residues starting at residue i in protein 1 and residue j in protein 2 with similarly shaped hydrophobicity, dark squares indicate dissimilar shapes. ($\boldsymbol{S}_{i=1,j=1}$ is the lower left corner in the figure.) As described in the text, $\boldsymbol{S}$ is exhaustively searched and all longest alignments with up to possibly *GapMax* gaps, whose squares (average path distance, *APD*) pass a user-defined average similarity cutoff *C,* are kept in a list (colored arrows). The alignment of this list with the closest absolute shape (lowest *RMSD*) is defined as the optimal match (Step 5). An Optimal Path Score (*OPS*), defined by Equation 4, is assigned to the alignment and its significance is computed with respect to the score distribution of random alignments of identical length (Step 6). Note that the example alignment, while a reasonable visual match, is only marginally significant with respect to random alignments of identical length, due to its short length of 10 residues.

Figure 2. Empirically determined probability models for biological data. A. Distributions of Equation 4 scores for *HePCaT* alignments of length $L = 100$ were obtained from parameters $W =$ 5 residues, *GapMax* = 4 residues, $C = 0.4$. Random data was simulated for Kyte-Doolittle

hydropathy averaged over a 9-residue window ("Hydrophobicity"), *eScape* predicted native state position-specific ΔG ("Thermodynamic Stability") and predicted codon-specific mRNA translation efficiency ("*tAI* Value"). Binned data in each case was reasonably fit to the Inverse Chi-Squared probability distribution function (PDF), as described in Methods and tabulated in Table S1. B. Parameters ν and σ^2 for the PDF were observed to vary smoothly as a function of *HePCaT* alignment length, allowing the parameters, and thus alignment significance, to be analytically estimated for any alignment length using Equations 6 and 7 and parameters in Table 1. Discrete best-fit parameters for ν and σ^2 are given in Table SI. Equations for displayed best-fit curves are as follows: y = 0.497609x (Hydrophobicity, ν), y = 0.740239x (Thermodynamic Stability, ν), y = 0.455863x (*E. coli* tAI Value, ν), y = 0.160379 – 1.04167 ln(x + 38.9045) (Hydrophobicity, σ^2), y = 9.11826 - 1.23341 ln(x + 45.3881) (Thermodynamic Stability, σ^2), y = -2.96918 – 0.961557 ln(x + 27.6082) (*E. coli* tAI Value, σ^2).

Figure 3. Most significant similarities in the human proteome to the Kyte-Doolittle hydropathy profile of adenosine receptor A2a. Pairwise *HePCaT* alignments are shown for A2a (black, gi|5921992) and the top nine most significant nonredundant hits in the human proteome. Blue color indicates known seven transmembrane proteins as annotated by the GPCR database, red mostly indicates hits to the tail region of A2a. The hits are shown from top to bottom in order of most to least significance: hematological and neurological expressed protein-like 1 (gi|21700763, $p = 4.5 \times 10^{-6}$), ephrin-A4 isoform a precursor (gi|4885197, $p = 8.3 \times 10^{-5}$), NSFL1 cofactor p47 isoform a (gi|20149635, $p = 9.8 \times 10^{-5}$), metallothionein-1E (gi|83367075, $p = 1.1 \times 10^{-4}$), taste receptor type 2 member 19 (gi|28882035, $p = 4.4 \times 10^{-4}$), B- and T-lymphocyte attenuator isoform 1 precursor (gi|145580621, $p = 5.7 \times 10^{-4}$), WD-repeat domain-containing protein 83 (gi|153791298, $p = 6.8 \times 10^{-4}$), dual specificity protein phosphatase 26 (gi|13128968, $p = 8.2 \times 10^{-4}$), adenosine receptor A2b (gi|4501951, $p = 9.1 \times 10^{-4}$). Thick lines indicate the optimal *HePCaT* alignment to A2a, and thin lines indicate unaligned positions.

Figure 4. Mammalian homologs exhibit greater thermodynamic similarity to the human N-terminal domain of glucocorticoid receptor than do non-mammailian homologs. The human protein (gi|121069, residues 1-414) is shown in black. Known AF1 (gray box) and scaffold (white box) functional regions are indicated above. Pairwise aligned positions of other homologs are shown below the human protein, in order of estimated significance of the match: 1. *H. sapiens* (gi|239758, residues 1-394, $p = 0$, exact match), 2. *H. sapiens* (gi|324021679, 1-99, $p = 0$, exact match), 3. *B. taurus* (gi|74354555, 1-418, $p = 1.0 \times 10^{-51}$), 4. *R. norvegicus* (gi|1189883, 1-433, $p = 0.33 \times 10^{-34}$), 5. *S. labiatus* (gi|121222567, 1-315, $p = 0.24 \times 10^{-30}$), 6. *R. norvegicus* (gi|152003264, 1-419, $p = 0.24 \times 10^{-30}$), 7. *B. taurus* (gi|38639409, 1-220, $p = 0.23 \times 10^{-19}$), 8. *R. norvegicus* (gi|56325, 1-433, $p = 1 \times 10^{-19}$), 9. *O. cuniculus* (gi|126723281, 1-409, p = 1.8×10^{-11}), 10. *O. anatinus* (gi|149632435, 1-412, $p = 1.6 \times 10^{-7}$), 11. *H. sapiens* (gi|221043882, 1-

17, p = 2.3x10^{-3}), 12. *A. carolinensis* (gi|327285250, 1-410, p = 2.8x10^{-3}), 13. *X. tropicalis* (gi|62858859, 1-415, p = 4.7x10^{-3}), 14. *C. carpro* (gi|219936801, 1-382, p = 2.1x10^{-2}), 15. *X. laevis* (gi|147905167, 1-413, p = 2.5x10^{-2}), 16. *O. latipes* (gi|253314476, 1-416, p = 0.10), 17. *T. guttata* (gi|224067332, 1-410, p = 0.25), 18. *M. domestica* (gi|126290524, 1-414, p = 0.38), 19. *S. trutta* (gi|57791246, 1-376, p = 0.43), 20. *P. promelas* (gi|66737265, 1-380, p = 0.55), 21. *C. carpro* (gi|156713894, 1-358, p = 0.59), 22. *C. pyrrhogaster* (gi|319412066, 1-408, p = 0.74), 23. *D. rerio* (gi|99028943, 1-382, p = 0.82). The dashed line indicates a *HePCaT* significance threshold of p = 0.01. Proteins from mammals are shown in red (generally above the significance threshold) and proteins from amphibians and fish are shown in blue (generally below the significance threshold). As indicated, the *A. carolinensis* protein exhibits significant thermodynamic similarity in its AF1 region but not the scaffold region, while the *X. tropicalis* protein exhibits significant similarity in the scaffold region but less so in AF1. Annotated partial transcripts are labeled, and known human isoforms with alternative translation start sites are marked with asterisks. Thick lines indicate the optimal *HePCaT* alignment to the query, thin lines indicate unaligned regions.

Figure 5. Translation efficiencies of homologous *E. coli* mRNA sequences do not appear to be conserved. The mRNA sequences of 377 pairs of *E. coli* proteins, each pair belonging to the same SCOP family, were optimally aligned using *HepCaT* based on the predicted translation efficiency *tAI* at each mRNA codon. The distribution of significance values for the set of optimal alignments is shown in dark circles; this distribution is not substantially different from a uniform distribution (dashed line). As a control, significances for the optimal alignments of Kyte-Doolittle hydropathy values for the same proteins exhibited a markedly skewed distribution (light squares), suggesting that hydrophobicity is more conserved than *tAI* for these proteins.

Figure 6. Predicted remote homology between *C. muridarum TC0624* and colicin pore-forming domain based on significant *HePCaT* similarity. A. Significant similarity between hydropathy of *TC0624* and *E. coli* colicin A (SCOP domain d1cola_).(4, 10) The likelihood of obtaining this match by chance is p = 5.5 x 10^{-6}. The blue circles indicate Jpred3 (12) predicted helical secondary structure of TC0624, the pink circles indicate the actual helical secondary structure of d1cola_ domain. Good correspondence between the type and location of secondary structure is observed. **B.** Tertiary structure location of the hydrophobic similarity (left) and the sequence similarity (right) matches between *TC0624* and colicin fold. In both molecular cartoons, helices are colored red, strands yellow, and loops green. Locations of a match between *TC0624* and colicin are colored blue. The left figure is based on d1cola_ and the right figure is based on the homolog d1a87a_ observed in the HHPred (47) hidden markov model match. This extensive structure, sequence, and chemical similarity between *TC0624* and colcin suggests the medically important hypothesis that *TC0624* is also a pore-forming protein that facilitates chlamydia survival.

Figure S1. Multiple sequence alignment of full length human glucocorticoid receptor and homologs with high sequence conservation exhibits less conservation in the N-terminal domain AF1 and scaffold regions. Non-redundant sequence homologs were collected using a *BLAST* search of the NCBI *nr* database as described in the main text. The alignment was generated using *PROMALS3D* with default parameters. The functionally important AF1 and scaffold regions of the N-terminal domain, referred to in Figure 4 of the main text, are indicated. Columns in these regions are less conserved than the rest of the alignment, as indicated by the colored annotations and the consensus summary, as annotated by *CHROMA.*

Figure 1. Overview of the Horizontal Protein Comparison Tool (*HePCaT*) algorithm.

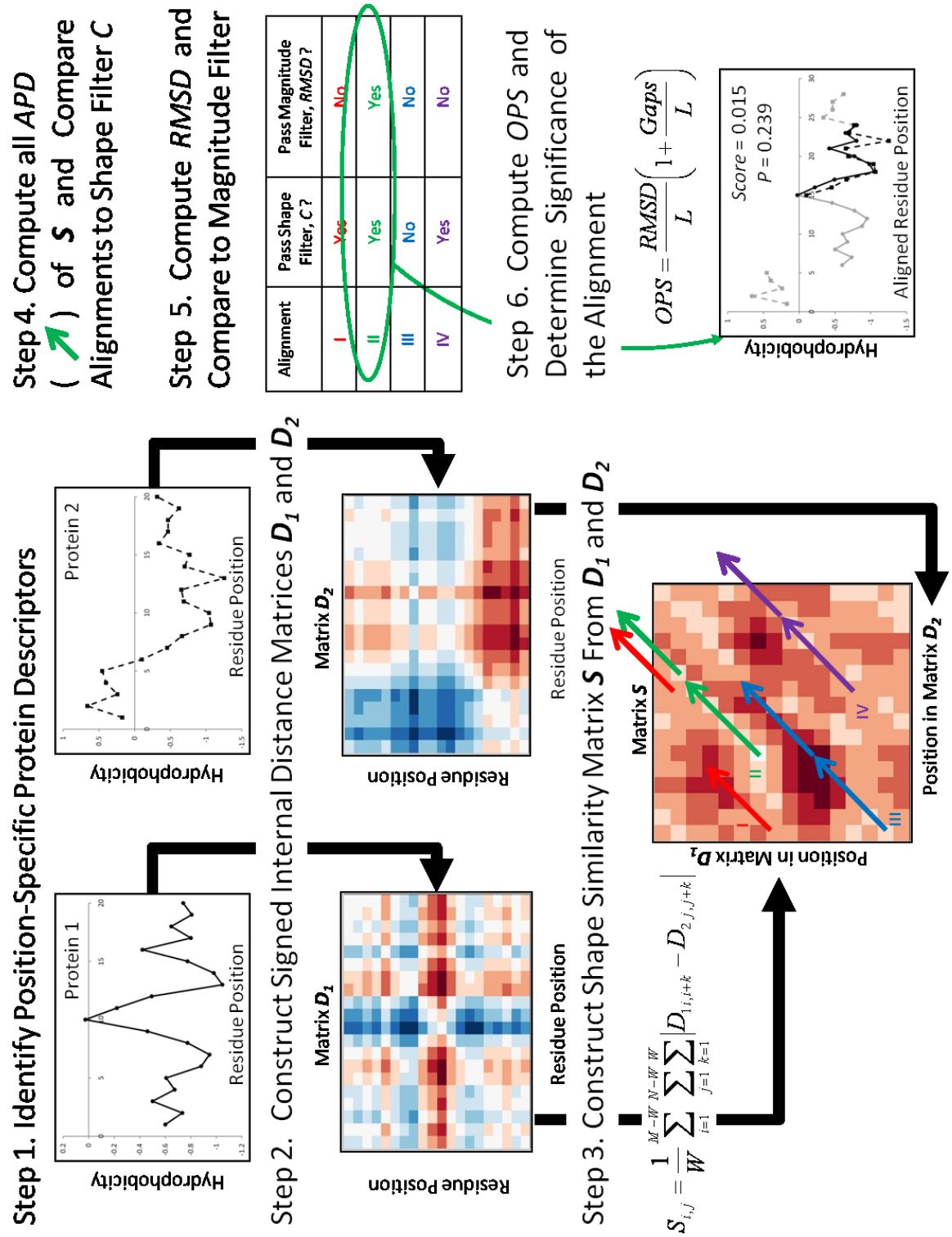


Figure 2. Empirically determined probability models for biological data.

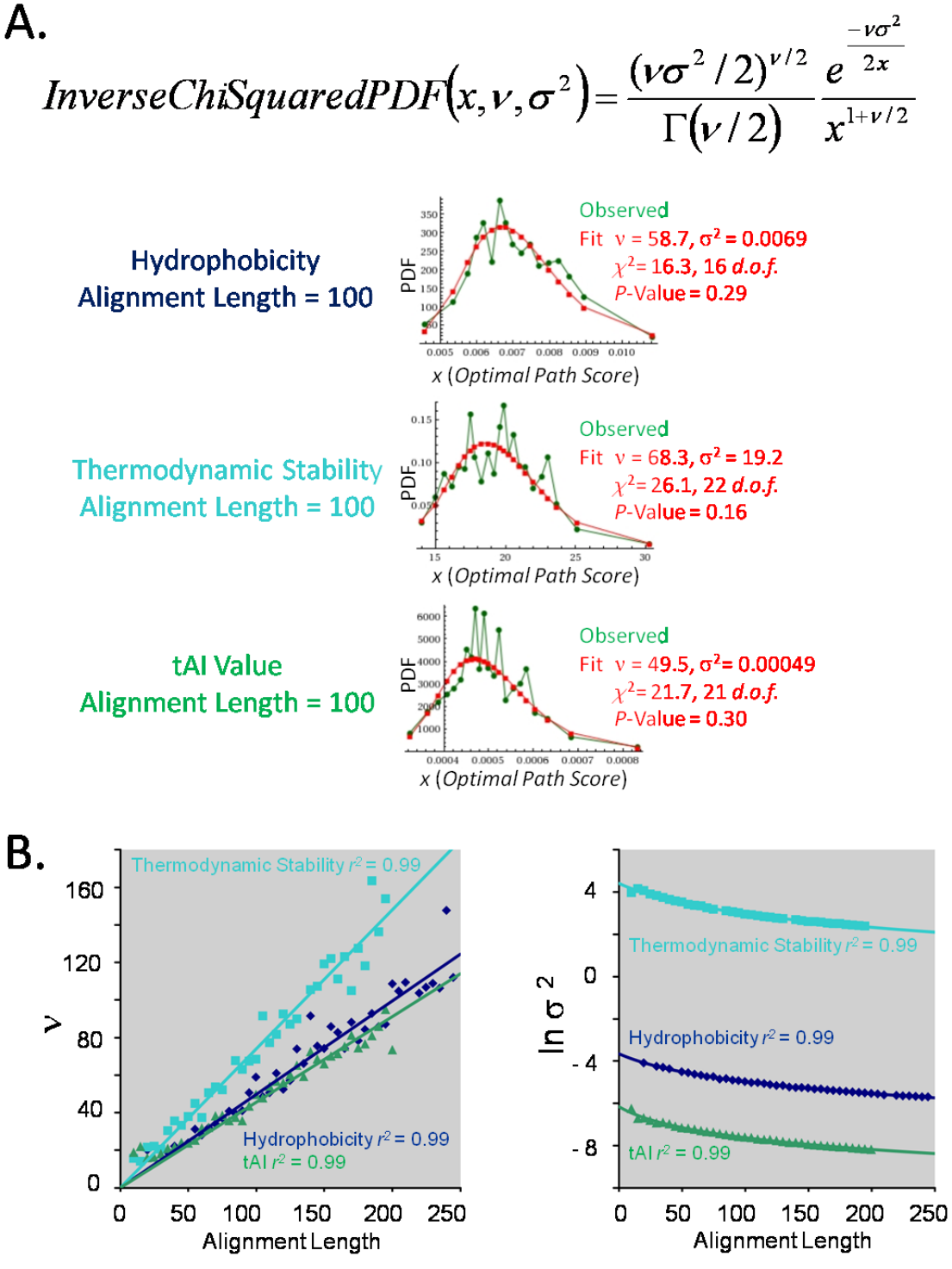


Figure 3. Most significant similarities in the human proteome to the Kyte-Doolittle hydropathy profile of adenosine receptor A2a.

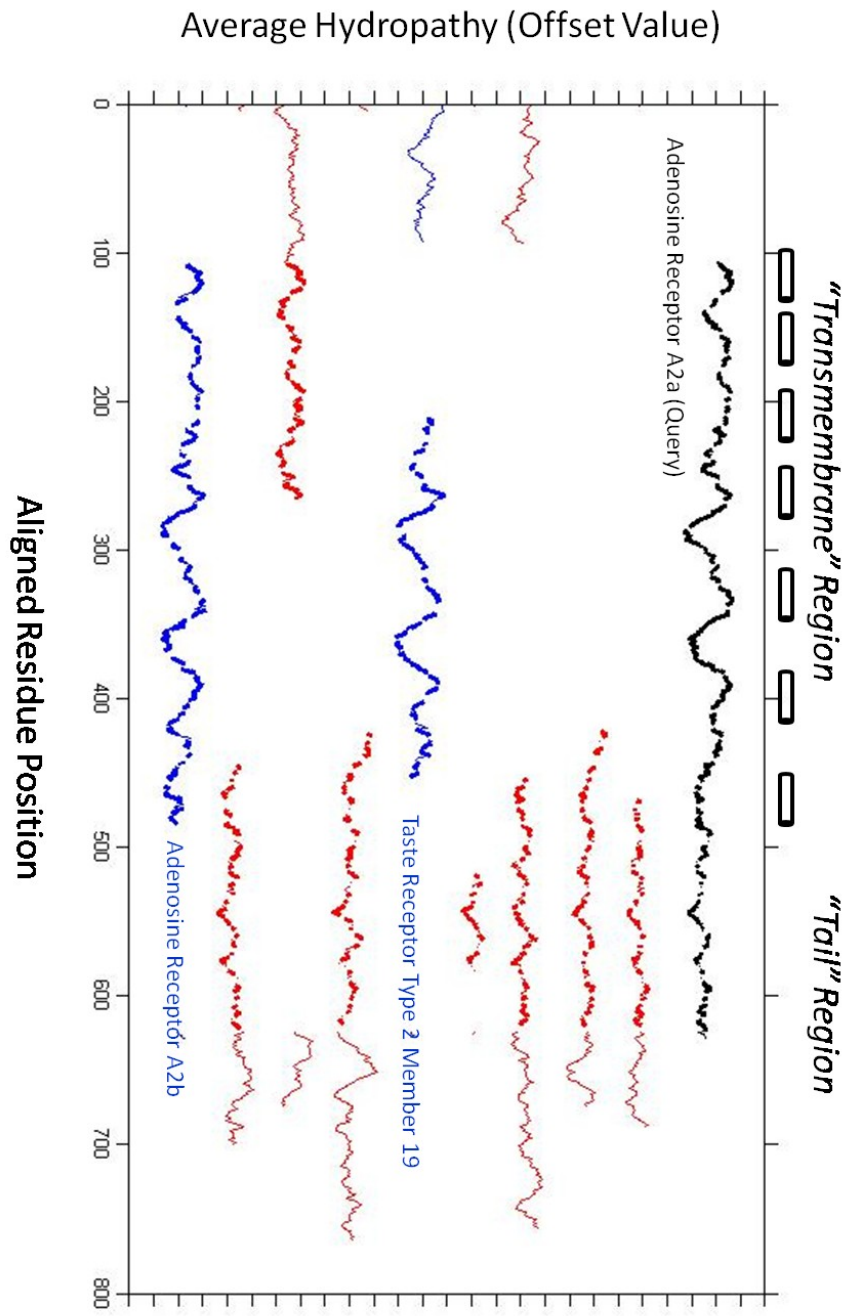


Figure 4. Mammalian homologs exhibit greater thermodynamic similarity to the human N-terminal domain of glucocorticoid receptor than do non-mammailian homologs.

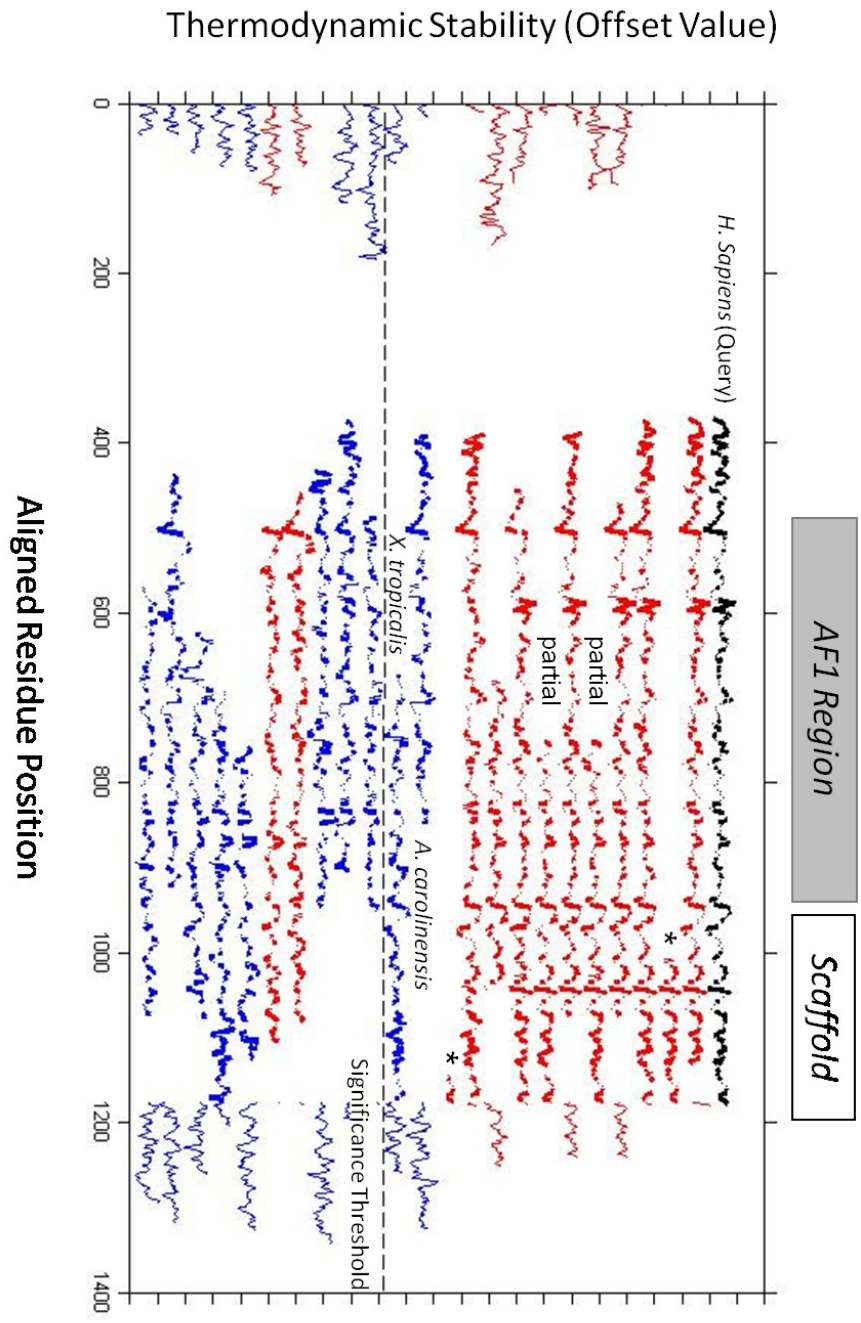


Figure 5. Translation efficiencies of homologous *E. coli* proteins do not appear to be conserved.

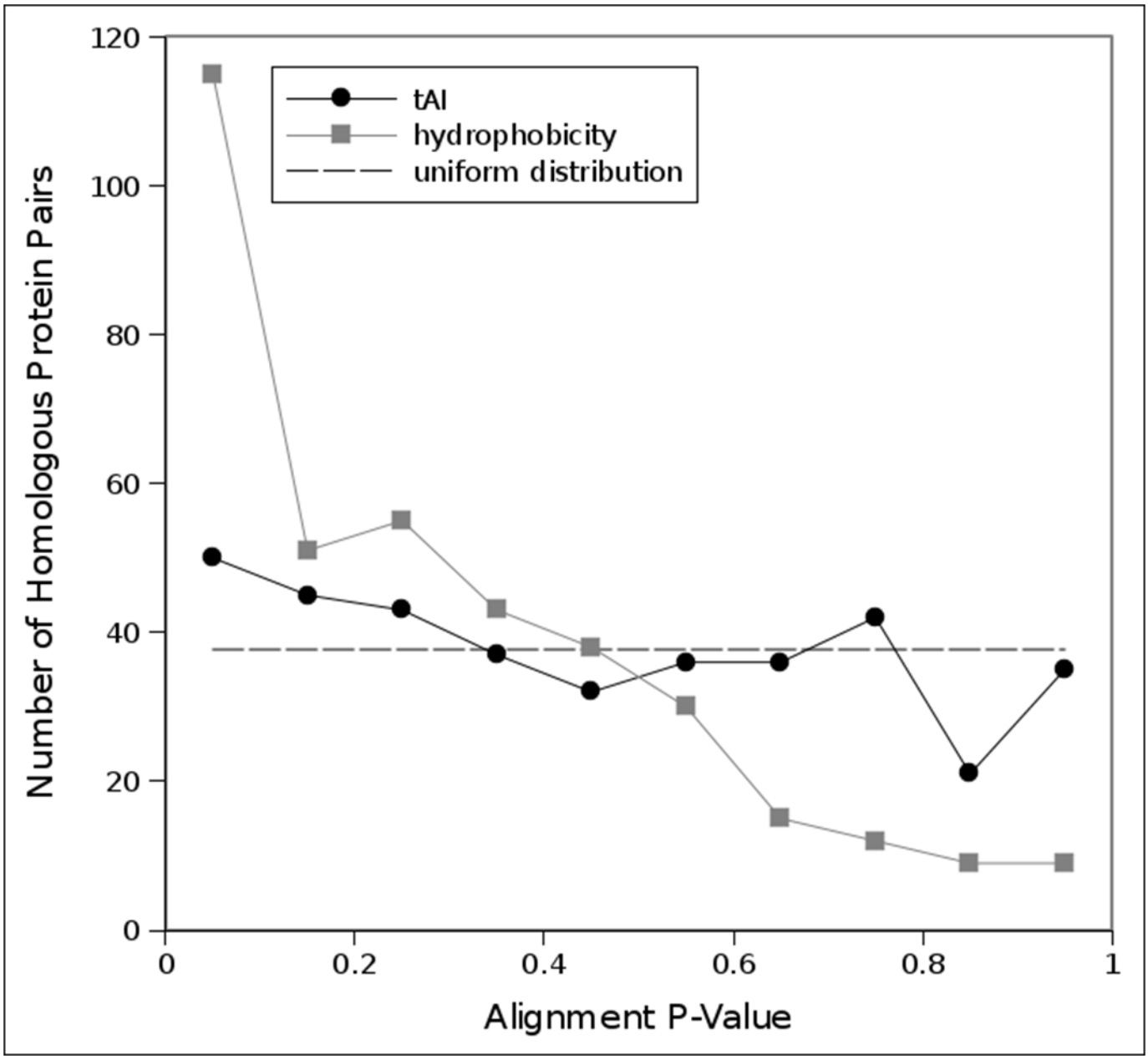
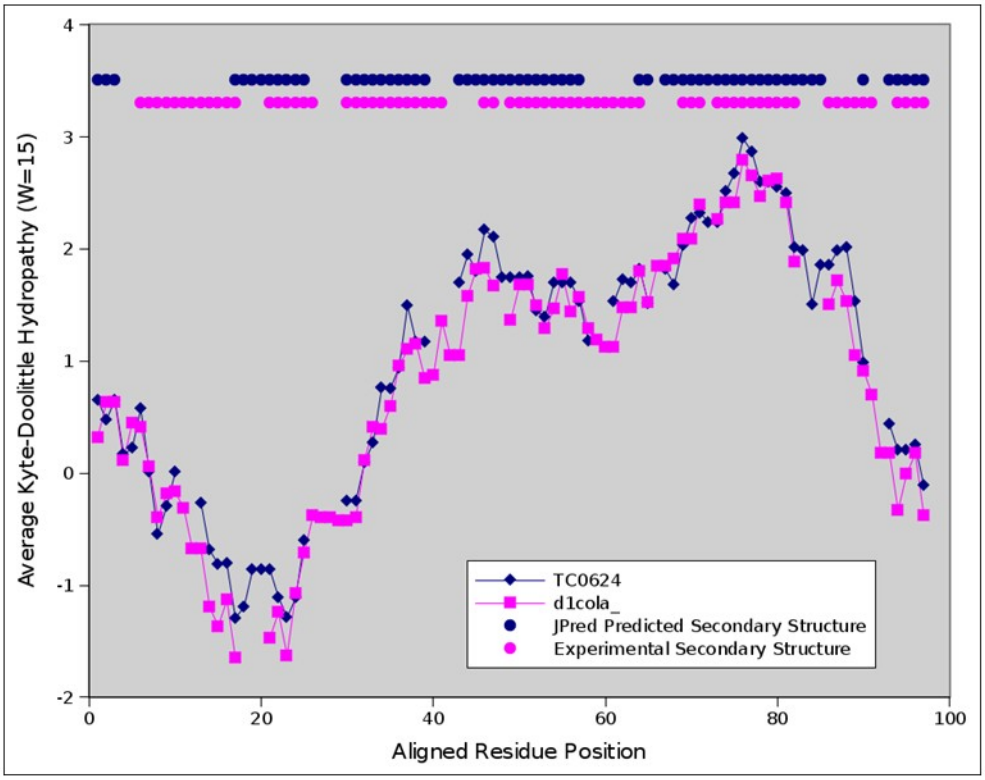


Figure 6. Predicted remote homology between *C. muridarum TC0624* and colicin pore-forming domain based on significant *HePCaT* similarity.

A.



B.

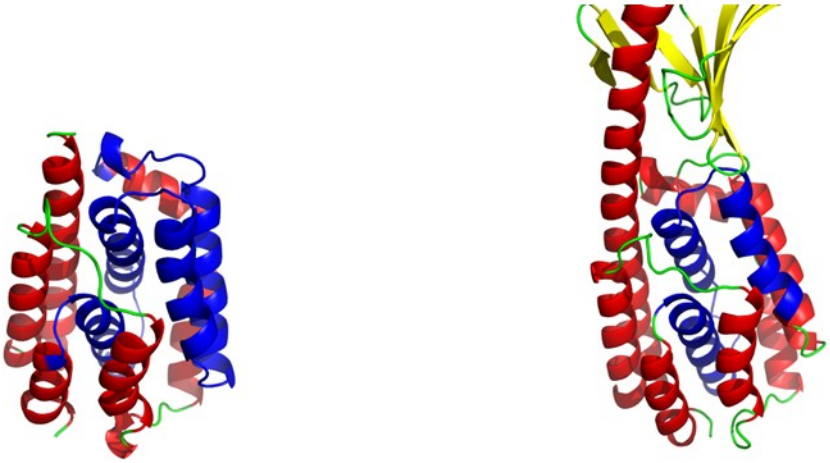
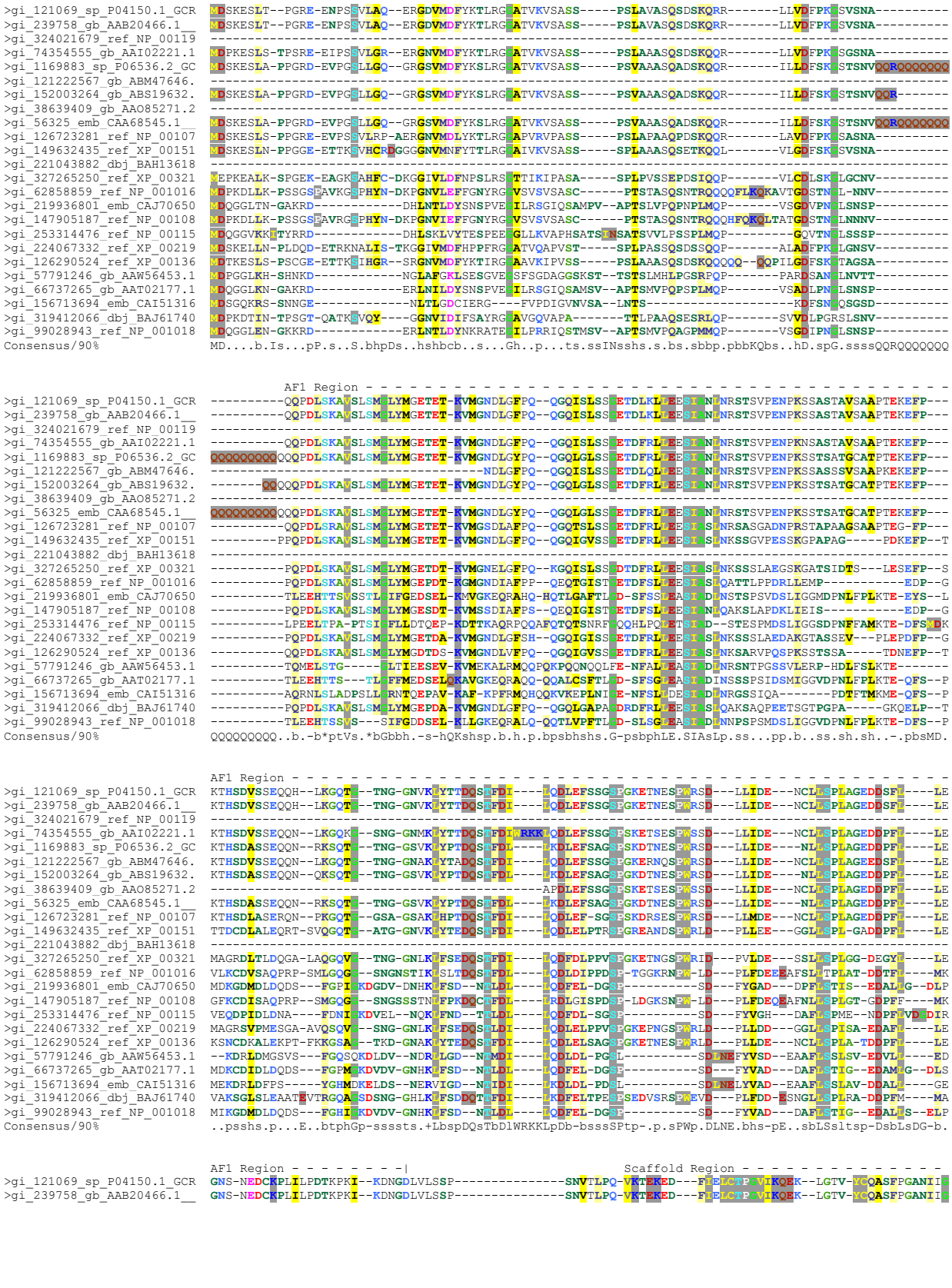
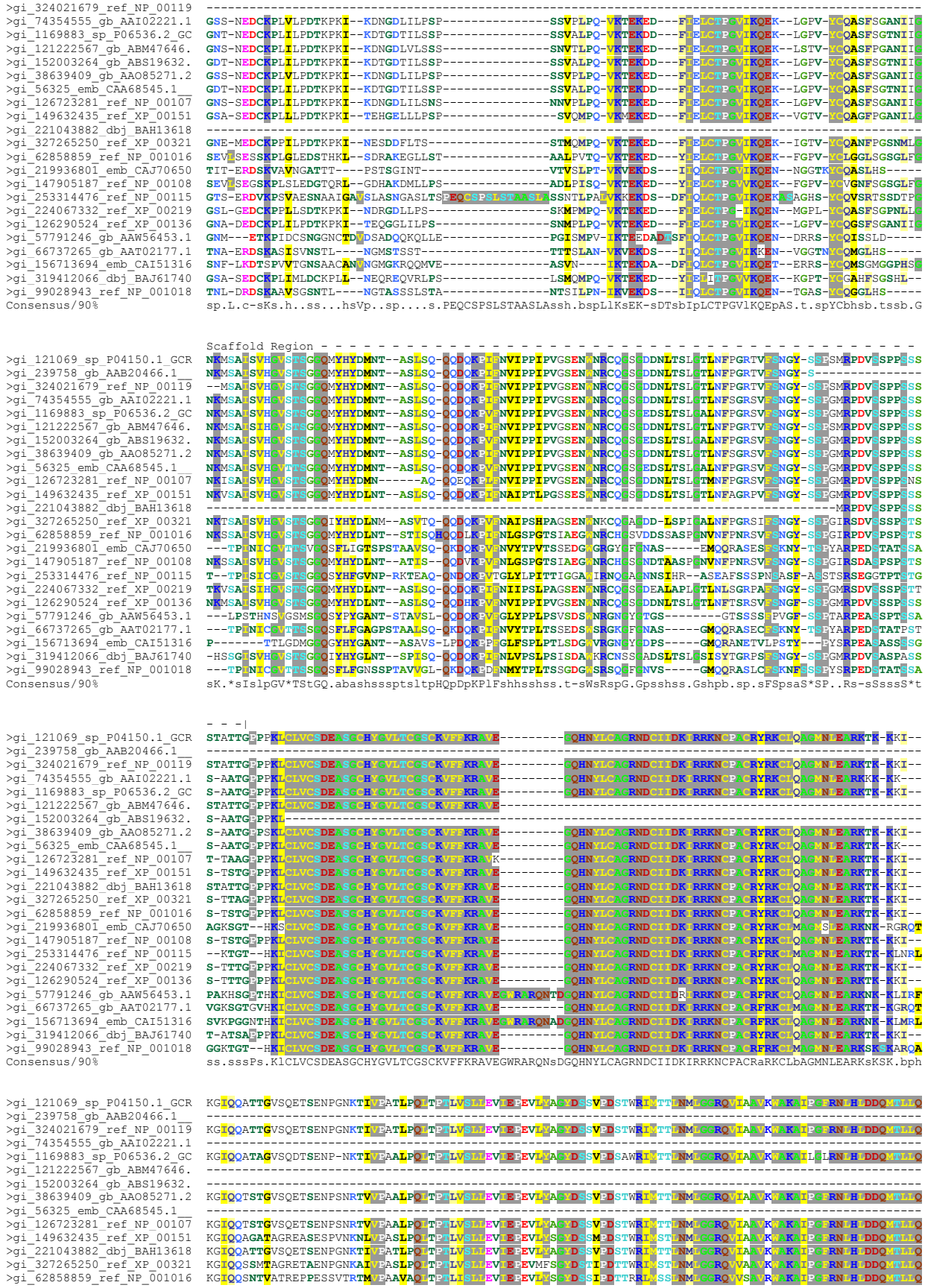


Figure S1. Multiple sequence alignment of full length human glucocorticoid receptor and homologs with high sequence conservation exhibits less conservation in the N-terminal domain AF1 and scaffold regions.





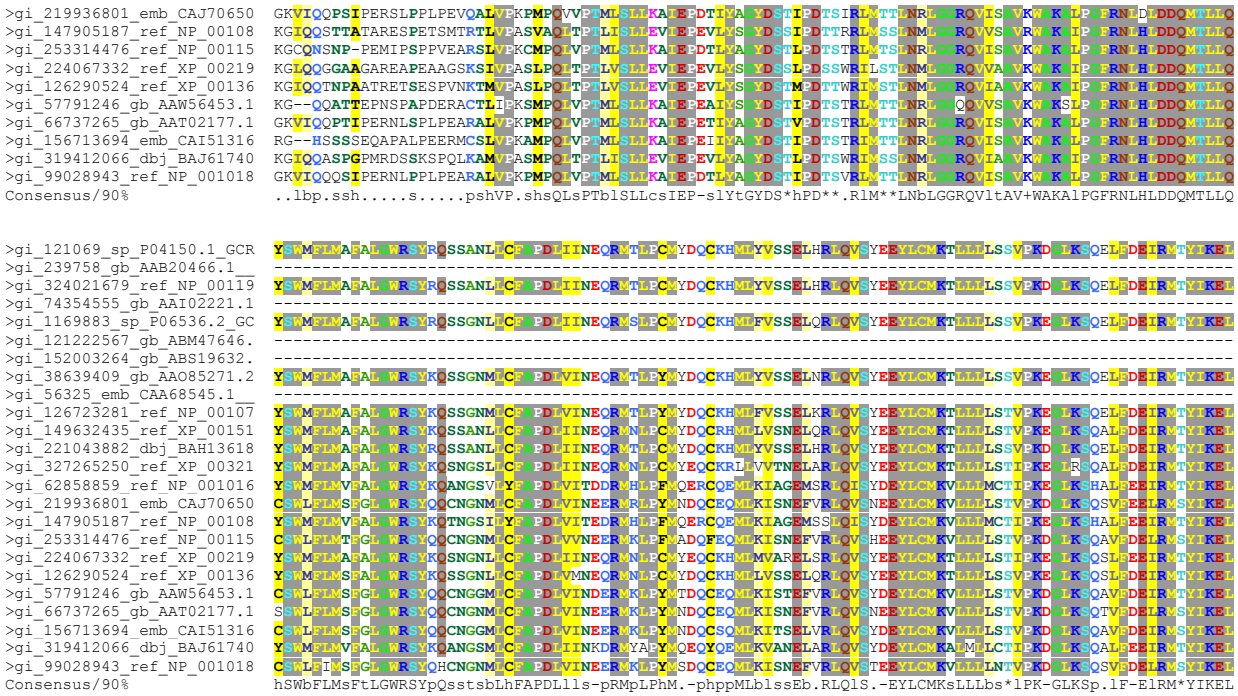

>gi_219936801_emb_CAJ70650
>gi_147905187_ref_NP_00108
>gi_253314476_ref_NP_00115
>gi_224067332_ref_XP_00219
>gi_126290524_ref_XP_00136
>gi_57791246_gb_AAW56453.1
>gi_66737265_gb_AAT02177.1
>gi_156713694_emb_CAI51316
>gi_319412066_dbj_BAJ61740
>gi_99028943_ref_NP_001018
Consensus/90%
>gi_121069_sp_P04150.1_GCR
>gi_239758_gb_AAB20466.1__
>gi_324021679_ref_NP_00119
>gi_74354555_gb_AAI02221.1
>gi_1169883_sp_P06536.2_GC
>gi_121222567_gb_ABM47646.
>gi_152003264_gb_ABS19632.
>gi_38639409_gb_AAO85271.2
>gi_56325_emb_CAA68545.1__
>gi_126723281_ref_NP_00107
>gi_149632435_ref_XP_00151
>gi_221043882_dbj_BAH13618
>gi_327265250_ref_XP_00321
>gi_62858859_ref_NP_001016
>gi_219936801_emb_CAJ70650
>gi_147905187_ref_NP_00108
>gi_253314476_ref_NP_00115
>gi_224067332_ref_XP_00219
>gi_126290524_ref_XP_00136
>gi_57791246_gb_AAW56453.1
>gi_66737265_gb_AAT02177.1
>gi_156713694_emb_CAI51316
>gi_319412066_dbj_BAJ61740
>gi_99028943_ref_NP_001018
Consensus/90%

Table S1. Goodness of fit statistics between Inverse Chi Square probability distribution function and *OPS* score distributions of various length optimal *HePCaT* alignments of random amino acid sequences. Blank rows indicate that the null hypothesis (*i.e.* that the random distribution of *OPS* scores was drawn from an underlying inverse chi square distribution) was rejected at the $p < 0.05$ level.

"Hydrophobicity" Kyte-Doolittle Hydropathy, averaged over 9 residues *W* = 5 residues *GapMax* = 4 residues *C* = 0.4						
Alignment Length	ν	$ln\ \sigma^2$	χ^2	***d.o.f.***	***P-Value***	***N***
20	20.030	-4.092	19.4	20	0.37	433
25						
30	20.444	-4.266	2.3	9	0.94	205
35	18.771	-4.305	9.8	15	0.71	322
40	22.152	-4.371	16.8	17	0.33	365
45						
50	23.895	-4.507	10.2	14	0.60	309
55	31.086	-4.556	23.7	17	0.07	368
60	27.883	-4.634	17.9	19	0.39	414
65	31.871	-4.675	9.1	17	0.87	379
70	34.017	-4.751	11.3	15	0.58	339
75	37.144	-4.752	16.0	19	0.52	405
80	40.667	-4.860	15.5	19	0.56	419

85	39.468	-4.851	19.2	17	0.21	374
90	40.866	-4.903	15.1	16	0.37	343
95	50.460	-4.935	19.0	18	0.27	386
100	58.710	-4.974	16.3	16	0.29	352
105	48.502	-5.033	15.0	15	0.31	329
110	50.481	-5.038	5.4	11	0.80	254
115	60.850	-5.074	6.9	14	0.86	315
120	52.309	-5.114	8.6	12	0.57	267
125	56.929	-5.160	7.4	13	0.76	295
130	73.921	-5.170	11.6	12	0.31	279
135	66.086	-5.231	3.7	13	0.98	282
140	91.441	-5.262	8.4	11	0.50	251
145	75.360	-5.265	4.6	12	0.92	276
150	74.003	-5.289	5.2	13	0.92	296
155						
160	82.535	-5.341	8.7	14	0.73	308
165	74.069	-5.378	7.9	15	0.85	331
170	87.990	-5.403	12.0	14	0.45	319
175	78.128	-5.437	19.1	17	0.21	362
180	84.227	-5.449	22.2	17	0.10	360
185	92.662	-5.472	9.8	15	0.71	332
190	85.812	-5.493	12.0	16	0.61	343
195	86.967	-5.531	12.7	16	0.55	344

200	108.592	-5.540	12.5	14	0.41	319
205	104.753	-5.565	13.1	15	0.44	332
210	109.308	-5.603	9.8	14	0.64	317
215						
220	103.593	-5.631	11.9	12	0.29	262
225	106.655	-5.651	9.2	12	0.51	260
230	108.842	-5.658	5.1	9	0.65	213
235	106.144	-5.687	9.1	9	0.25	203
240	147.619	-5.705	6.2	9	0.52	201
245	111.964	-5.717	4.7	7	0.45	173

"Thermodynamic Stability" eScape Native State ΔG Value (cal/mol @ 25 °C) *W* = 5 residues *GapMax* = 4 residues *C* = 0.4						
Alignment Length	ν	$ln\ \sigma^2$	χ^2	***d.o.f.***	***P-Value***	***N***
10	15.543	3.966	2.8	6	0.59	153
15	13.934	4.162	27.5	34	0.69	717
20	21.620	4.058	2.9	6	0.58	145
25	22.057	3.906	9.9	7	0.08	167
30	20.573	3.820	23.3	28	0.11	391
35	30.307	3.740	10.6	22	0.96	472

40	35.467	3.639	4.1	10	0.85	227
45	33.262	3.567	9.1	13	0.61	298
50	37.757	3.521	25.6	18	0.06	385
55	44.858	3.417	14.6	18	0.55	380
60	37.390	3.366	18.5	16	0.18	340
65	50.430	3.328	17.8	18	0.34	386
70	53.544	3.242	16.3	21	0.64	441
75	52.196	3.172	15.0	19	0.59	404
80						
85	67.643	3.112	8.7	19	0.95	412
90	62.877	3.048	16.9	21	0.60	446
95	67.465	3.009	28.2	23	0.13	481
100	68.351	2.954	26.1	22	0.16	473
105	91.331	2.909	29.0	23	0.11	488
110	77.172	2.878	13.7	23	0.88	489
115	81.511	2.824	17.5	25	0.79	523
120	92.397	2.804	16.0	25	0.86	532
125	86.950	2.765	30.2	30	0.35	620
130	89.843	2.731	22.9	26	0.53	543
135						
140	105.431	2.675	22.1	26	0.57	546
145	107.312	2.641	27.5	25	0.23	522
150	119.116	2.605	22.2	24	0.45	515

155	121.909	2.590	25.1	21	0.16	452
160	111.151	2.558	12.7	19	0.76	401
165	122.935	2.516	20.7	17	0.15	377
170	104.958	2.491	17.7	16	0.22	352
175	127.507	2.493	20.1	14	0.07	305
180	118.001	2.447	14.0	14	0.30	307
185	163.377	2.421	13.4	12	0.20	262
190	136.289	2.412	8.7	8	0.19	181
195	153.717	2.377	7.4	7	0.20	168

"Translation Efficiency" *E. coli* tAI Value, averaged over 9 codons *W* = 5 residues *GapMax* = 4 residues *C* = 0.4						
Alignment Length	ν	$\ln \sigma^2$	χ^2	***d.o.f.***	***P-Value***	***N***
10	18.724	-6.278	6.2	6	0.18	156
15	21.753	-6.739	1.0	15	0.80	126
20	15.883	-6.714	15.0	18	0.52	398
25	16.017	-6.816	15.5	21	0.69	446
30	17.772	-6.940	3.3	9	0.85	206
35	20.707	-6.924	13.4	14	0.34	312
40	21.598	-6.984	15.9	18	0.46	393

45	23.766	-7.111	10.6	17	0.78	362
50	23.655	-7.172	12.2	14	0.43	304
55	25.040	-7.209	13.6	14	0.43	318
60	31.666	-7.242	11.4	17	0.72	375
65	31.851	-7.307	21.6	19	0.20	413
70	37.965	-7.365	12.8	17	0.61	364
75	38.406	-7.391	16.4	17	0.36	378
80	35.419	-7.461	15.2	19	0.58	406
85	37.307	-7.508	15.9	19	0.53	414
90	35.574	-7.545	17.0	19	0.46	410
95	43.313	-7.587	14.2	21	0.77	440
100	49.518	-7.626	21.7	21	0.30	440
105	47.622	-7.651	10.7	21	0.93	448
110	53.770	-7.703	19.4	21	0.43	450
115	56.545	-7.745	24.5	24	0.32	504
120	55.624	-7.743	32.6	23	0.05	484
125	60.158	-7.791	19.4	20	0.37	438
130	65.094	-7.831	12.3	23	0.93	499
135	59.084	-7.867	14.9	24	0.87	508
140	72.233	-7.885	22.9	23	0.35	480
145	68.343	-7.908	16.3	21	0.63	455
150	65.821	-7.960	18.0	23	0.65	495
155	70.033	-7.983	15.7	20	0.62	427

160	71.137	-8.008	16.1	18	0.44	397
165	75.067	-8.031	12.0	18	0.75	389
170	80.949	-8.066	20.7	17	0.15	374
175	74.439	-8.094	13.8	14	0.31	308
180	76.417	-8.093	10.6	14	0.56	307
185	80.811	-8.139	6.5	11	0.69	246
190	85.824	-8.150	9.6	10	0.29	229
195	94.885	-8.170	5.9	7	0.31	175
200	73.380	-8.208	4.6	7	0.47	163