Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Building multiple sequence alignments with a flavor of HSSP alignments

Roberto Hiroshi Higa, Sergio Aparecido Braga da Cruz, Paula Regina Kuser, Michel Eduardo Beleza Yamagishi, Renato Fileto, Stanley Robson de Medeiros Oliveira, Ivan Mazoni, Edgard Henrique dos Santos, Adauto Luiz Mancini and Goran Neshich

Núcleo de Bioinformática,
Centro Nacional de Pesquisa em Informática Agropecuária,
Empresa Brasileira de Pesquisa Agropecuária, Campinas, SP, Brasil
Corresponding author: G. Neshich
E-mail: neshich@cbi.cnptia.embrapa.br

Genet. Mol. Res. 5 (1): 127-137 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. Homology-derived secondary structure of proteins (HSSP) is a well-known database of multiple sequence alignments (MSAs) which merges information of protein sequences and their three-dimensional structures. It is available for all proteins whose structure is deposited in the PDB. It is also used by STING and ^{Java}Protein Dossier to calculate and present relative entropy as a measure of the degree of conservation for each residue of proteins whose structure has been solved and deposited in the PDB. However, if the STING and ^{Java}Protein Dossier are to provide support for analysis of protein structures modeled in computers or being experimentally solved but not yet deposited in the PDB, then we need a new method for building alignments having a flavor[1] of HSSP alignments (myMSAr). The present study describes a new method and its corresponding databank (SH_2Q^S - database of sequences homologue to the query [structure-having] sequence). Our main interest in making myMSAr was to measure the degree of residue conservation for a given query sequence, regardless of whether it has a corresponding structure deposited in the PDB. In this study, we com-

[1]We use the term flavor of HSSP to emphasize the fact that our procedure mimics the HSSP procedure for building multiple sequence alignments and that the relative entropy values reported by myMSAr are to be as close as possible to the ones reported by the HSSP.

pare the measurement of residue conservation provided by corresponding alignments produced by HSSP and SH_2Q^S. As a case study, we also present two biologically relevant examples, the first one highlighting the equivalence of analysis of the degree of residue conservation by using HSSP or SH_2Q^S alignments, and the second one presenting the degree of residue conservation for a structure modeled in a computer, which , as a consequence, does not have an alignment reported by HSSP.

Key words: Multiple sequence alignment, Residue conservation, MSA, HSSP, SH_2Q^S, Relative entropy

INTRODUCTION

STING and [Java]Protein Dossier (Neshich et al., 2003, 2004, 2005a,b) are web-based tools that support the study of the complex interplay of protein sequence/structure/function. These tools are available for the analysis of proteins whose three-dimensional structure has been either solved or modeled. The STING database contains a number of physical-chemical parameters (currently more than 300 of them) shown on a residue by residue basis. In particular, the values of relative entropy[2] reported by homology-derived secondary structure of proteins (HSSP) (Sander and Schneider, 1991) are used as a measure of degree of residue conservation.

The HSSP is a derived database which merges information from three-dimensional protein structures and sequences of proteins. For each entry in the Protein Data Bank (PDB) (Berman et al., 2000), the database reports an alignment containing a list of putative homologues selected from Uniprot/Swiss-prot (Apweiler et al., 2004). These alignments are built by using an iterative position-weight dynamic programming method - MaxHom. A threshold value for structural homology (Sander and Schneider, 1991) is used to decide if a sequence is a putative homologue or not. HSSP has been used to study protein evolution, folding and design and, in particular, to define structurally meaningful sequence patterns as well as to analyze residue conservation in a structural context.

However, as HSSP is available only for proteins whose structure is deposited in the PDB, conservation could not be reported by the STING and [Java]Protein Dossier for protein structures either modeled by computational methods (homology modeling), or experimentally deciphered but not yet deposited in the PDB. Moreover, being pre-calculated, HSSP does not allow the user to customize the value of the threshold nor to edit the multiple sequence alignment (MSA).

To overcome these limitations, we developed an alternative method for searching the sequence database and then filtering the homologous sequences by using a threshold value and building an MSA, which is aimed at resembling the one reported by HSSP. STING and [Java]Protein Dossier use this method to calculate the degree of residue conservation for all protein sequences which are and are not contemplated by HSSP. This process has been implemented as a program - myMSAr - written in Perl (Wall et al., 1996) and is based on the well-known software packages Blast (Altschul et al., 1997) and ClustalW (Thompson et al., 1994). We also

[2]In the present study, we use the term relative entropy as used by Sander and Schneider (1991). Note that it is different from the Kullback-Leibler distance (Cover and Thomas, 1991), also known as relative entropy in the field of Theory of Information.

built a database of alignments (SH_2Q^S - database of sequences homologue to the query [structure-having] sequence) by using this process for all protein sequences in the PDB.

The present study describes the processing used for building SH_2Q^S as well as a comparison of the degree of residue conservation reported by SH_2Q^S and HSSP. The comparison of the profiles from two alignments is also presented.

METHODS

The processing for building multiple sequence alignment

Our method for building MSAs having a flavor of HSSP is based on the software packages Blast and ClustalW. The input consists of a query sequence and, optionally, the e-value (default = 0.1) used by Blast, the gap open penalty (default = 3.0) and the gap extension penalty (default = 0.1) used by ClustalW, and a minimum homology threshold value (default = 5). Blast is used for making a search in Uniprot/Swiss-prot protein sequence databank (Apweiler et al., 2004) using the query sequence provided as input. The similarity matrix used was BLOSUM60 and all other Blast options were set to their default values. For filtering the homologue sequences, we used the percentage of identity, provided by the Blast's heuristic local alignment, as a measurement of similarity. It was compared with a local alignment length-dependent threshold function T(L), reported by Sander and Schneider (1991),

$$T(L) = 290 \cdot 15L^{-0.562} \qquad \text{(Equation 1)}$$

where L stands for the local sequence alignment length. For building the MSA with HSSP flavor using ClustalW, all other options were set as disabled and only the parts of sequences aligned by Blast were used for building the MSA. In particular, the SH_2Q^S databank was built by running the process using all selectable parameters set to their default values. As our process does not have a step for pair-wise alignment refinement, as does MaxHom, we tried to compensate for this by using more restrictive parameters (lower Blast e-value and higher homology threshold). Figure 1 illustrates this processing.

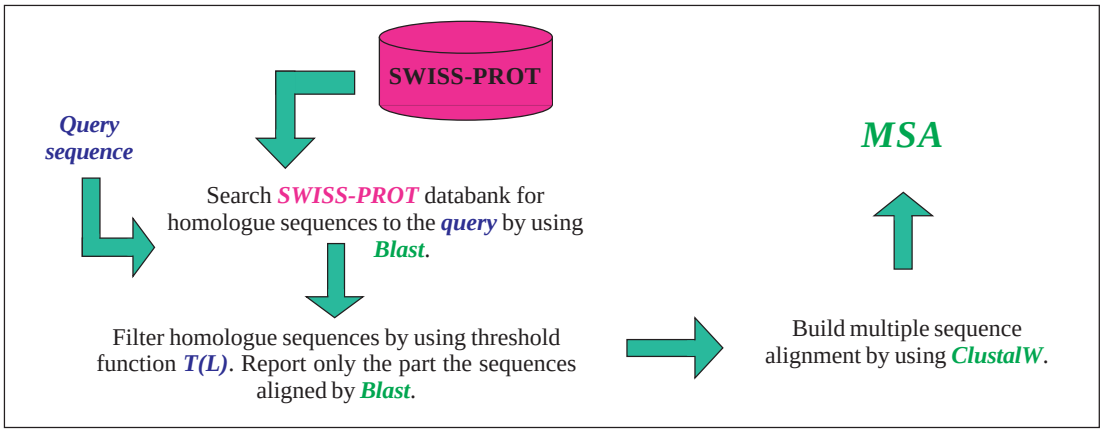


Figure 1. The processing implemented for building multiple sequence alignment (MSA) with the HSSP flavor. It is based on the use of the Blast and ClustalW tools and Uniprot/Swiss-Prot database, with a threshold function to select homologue sequences.

Relative entropy

Relative entropy, as defined by Sander and Schneider (1991), is used by HSSP to report the degree of conservation of each residue corresponding to an HSSP entry. It is calculated as follows: for each position of the MSA, the relative frequency of each of the 20 amino acids is calculated. The entropy at position "x" is then given by

$$S_x = -\sum_i p_i \log(p_i) \qquad \text{(Equation 2)}$$

where p_i is the relative frequency of amino acid "i" at position "x", $1 \le i \le 20$.

The relative entropy ("relent") at position "x" is given by the ratio of the entropy at that position and the entropy for the uniform distribution, which corresponds to the maximum entropy.

$$relent(x) = \frac{S_x}{S_{umif}} \qquad \text{(Equation 3)}$$

Measuring similarity

To measure the similarity between the values of relative entropy calculated by using the corresponding MSAs reported by the HSSP and SH_2Q^S, we used the cosine formed by the profiles of residue conservation. The cosine can be expressed as:

$$\cos ine = \frac{\langle X, Y \rangle}{\| X \| \| Y \|} \qquad \text{(Equation 4)}$$

where X is a vector and each element X_i represents a value for the relative entropy, calculated for the position "i" of the query sequence using the MSA reported by SH_2Q^S; Y is a vector and each element Y_i represents a value for the relative entropy, calculated for the position "i" of the query sequence using the MSA reported by the HSSP. As all values for the relative entropy are positive, cosine values vary from 0 to 1. Values close to 1 indicate high similarity.

Alternatively, we also wanted to determine to what extent a query sequence had residues with overly different values of relative entropy for corresponding HSSP and SH_2Q^S alignments. For this, we calculated for each query sequence the percentage of residues whose difference in relative entropy (calculated by using MSAs reported by HSSP and SH_2Q^S) was: a) higher than 30%; b) higher than 40%; c) higher than 50%, and d) higher than 60%. These values were chosen because we considered that for such a difference in the evaluation of the degree of conservation, one could certainly say that there is a significant difference between the two methods.

Dataset

To compare the results produced by HSSP with those produced by SH_2Q^S, we selected a dataset based on HSSP released in May, 2004. That HSSP release contains 23,868 HSSP files, corresponding to 36,144 sequences having an alignment (some HSSP files have alignments for more than one sequence as the PDB file has more than one chain). At the time we conducted the experiment, SH_2Q^S reported 51,075 sequences having an alignment, corresponding to the Uniprot/Swiss-Prot release 3.0/45.0 of October 25, 2004 and PDB release of December 6, 2004. The extra alignments at SH_2Q^S is due to two facts: a) for dimers, the HSSP reports only one alignment while SH_2Q^S reports one alignment for each sequence, and b) the HSSP does not accompany the dynamics of the PDB updates and, consequently, it does not include the MSAs for the most recent PDB entries.

From the original 36,144 sequences in the HSSP, there were 35,573 sequences for which SH_2Q^S reported alignments and 571 sequences for which it did not. This was due to the fact that the value of the parameters we used for building SH_2Q^S were more restrictive than those used by HSSP (consequently, myMSAr did not generate any MSA for some entry sequences). In addition, there were 2369 sequences for which we could not easily determine the cosine value for the relative entropy reported by HSSP and SH_2Q^S. This was due to the fact that HSSP does not report in query sequence those residues having missing atomic coordinates, while SH_2Q^S does. Although we could handle this situation by re-aligning the query sequence in the corresponding MSAs to identify the part reported by both methods, we decided not to consider these alignments in this study. Consequently, the final dataset consisted of 33,204 sequences having alignments reported by both HSSP and SH_2Q^S, and for which we could readily determine the cosine between the relative entropies of the respective alignments. All experiments were carried out with this dataset.

RESULTS

For a given query sequence, most of the time HSSP reported alignments containing more aligned sequences than the corresponding alignment produced by the SH_2Q^S. The alignments reported by HSSP had 3 to 3001 aligned sequences (mean equal to 290.5), whereas those reported by SH_2Q^S had 2 to 251 aligned sequences (mean equal to 83.1). This fact was a consequence of the values of the parameters used to generate the SH_2Q^S alignments, which were more restrictive than those used by HSSP.

Comparison of the relative entropy values calculated from HSSP and SH_2Q^S alignments

The calculated mean value for the measure of similarity between values for the relative entropy calculated by the two methods (the cosine value) was 0.9 with a standard deviation of 0.1. This value indicates that although the values of relative entropy calculated from the HSSP and SH_2Q^S alignments may not be exactly the same, they have very similar profiles of residue conservation. To illustrate how similar the relative entropy profiles are along the single sequence, Figure 2 shows an X-Y plot of the values for the relative entropy calculated from the alignments generated by HSSP and SH_2Q^S, corresponding to the PDB id 1i6d, chain A. This

PDB chain refers to the structure of the functional site domain of *Paracoccus denitrificans* cytochrome C552 in the reduced state and has been solved by NMR (Pritovsek et al., 2000). This protein was chosen here as a typical protein because the number of residues and sequences in SH_2Q^S alignment are very close to the mean values of our dataset.

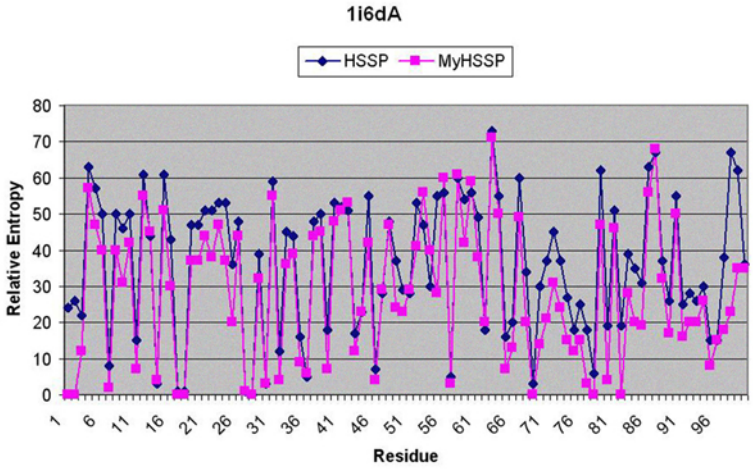


Figure 2. The relative entropy values reported along the protein sequence corresponding to PDB ID 1i6d: reported by the HSSP (blue) and SH_2Q^S [MyHSSP] (magenta) alignments.

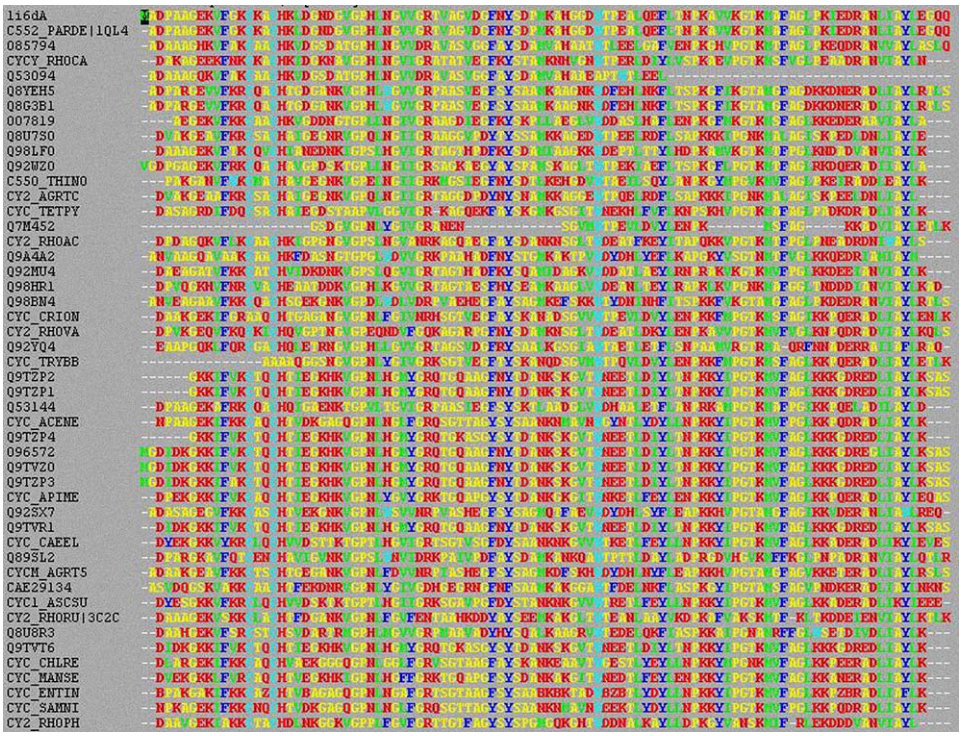
Both alignments, the HSSP and SH_2Q^S, used to calculate the relative entropies for the functional site domain of *Paracoccus denitrificans* cytochrome C552 are presented in Figure 3 for comparison of profiles.

An analysis of the errors in evaluating the degree of residue conservation

We also determined to what extent a query sequence comprises residues with overly different values of the relative entropy for the corresponding HSSP and SH_2Q^S alignments. We wanted to know, on average, the proportion of residues whose relative entropy calculated from the HSSP and SH_2Q^S alignments is so different that it could lead the reader/user to view the degree of residue conservation differently when using the values of one versus the other method.

Table 1 presents the number of alignments having at least one residue with the relative entropy value reported by HSSP being different from the one reported by SH_2Q^S. The table shows that the majority of the alignments showed at least one residue having overly different values for the relative entropy. However, the difference in relative entropy value per residue for most of them was small (being less than 30). In addition, we observed that the percentage of residues with a difference in relative entropy reported by the two methods does not exceed 12.3% of the residues of the corresponding query sequence on average and that it decreases as the level of difference in relative entropy considered increases.

A



B

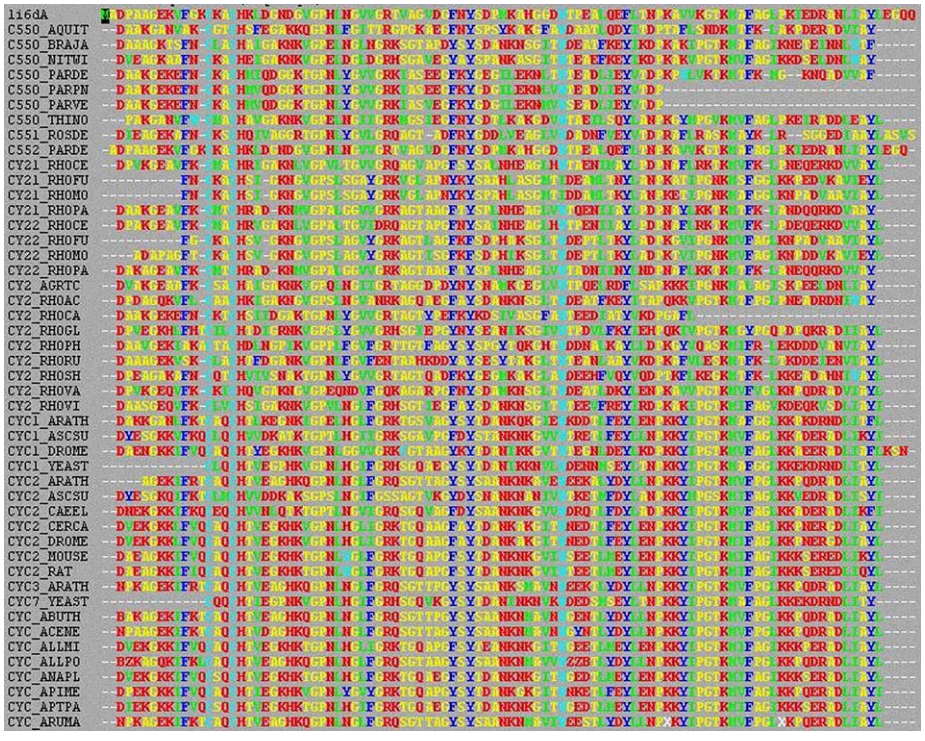


Figure 3. Alignments built by HSSP (A) and SH_2Q^S (B) for the functional site domain of the *Paracoccus denitrificans* cytochrome C552 (pdb id 1i6d, chain A). The first 47 sequences in each alignment are presented. In both alignments, gaps opened in the query sequence are not shown.

Table 1. Number of alignments having at least one residue showing a difference in value for the relative entropy reported by HSSP and SH_2Q^S. Four levels of difference in the relative entropy are considered: higher than 30%, higher than 40%, higher than 50%, and higher than 60%.

Difference in relative entropy	>30%	>40%	>50%	>60%
Number of alignments having at least one residue with overly different values of relative entropy reported by HSSP and SH_2Q^S	28312	23130	16719	10124
Average percentage, by alignment, of the residues having overly different values of relative entropy reported by HSSP and SH_2Q^S	12.31%	7.22%	4.41%	2.78%

HSSP = homology-derived secondary structure of proteins; SH_2Q^2 = sequences homologue to the query [structure-having] sequence.

Case study 1: HSSP/SH_2Q^S comparative analysis of conservation for residues of DNA polymerase I

We used DNA polymerase to validate the results reported by SH_2Q^S in comparison to those reported by HSSP. DNA polymerases contain an active site that may be structurally superimposed among many family members and is highly conserved in sequence. This protein has a central role in the process of transmission of genetic information over generations. Its overall folding contains three distinct sub-domains - palm, fingers and thumb (Beese et al., 1993). They share two conserved regions: motif A and motif C, both located within the palm sub-domain. Also, the primary sequence of various DNA polymerase active sites is exceptionally conserved, suggesting that motif A evolved slowly. In particular, the sequence DYSQIELR in motif A is conserved through polymerases from organisms separated by many million years of evolution, including *Thermus aquaticus*, *Chlamydia trachomatis* and *Escherichia coli*. Patel and Loeb (2000) performed site-directed mutagenesis experiments which indicated that all residues of motif A, except Asp:610, are mutable while preserving wild-type activity. Earlier, we analyzed the residues belonging to the active site by means of structure attributes from STING_DB (Neshich et al., 2004).

Here, we present the set of conserved residues for this protein obtained by using the relative entropy value for alignments built by HSSP and SH_2Q^S. For this protein, the measure of similarity of the relative entropy obtained by the two methods (cosine value) was 0.8, a little lower than the average value we found for the whole set of the HSSP/SH_2Q^S comparable alignments (0.9). Consequently, one can rush to the conclusion that the presented example can be considered as an unfavorable case in the use of SH_2Q^S. However, we performed an experiment to see how two sets of values for relative entropy (obtained from HSSP and SH_2Q^S) behave in terms of selecting a very similar subset of amino acids (say those with the highest conservation). By setting appropriately the selection parameters in ^{Java}Protein Dossier (the relative entropy value used as a threshold above which all residues were discarded from the view) first for the SH_2Q^S-derived relative entropy and then for the HSSP-derived relative entropy values, it was possible to select a very similar set of conserved residues. This is illustrated in

Figure 4. The fact that we have to use a relative entropy value of 8 as a threshold when selecting residues according to the SH_2Q^S alignments (instead of 0 which is used in the case of HSSP-derived values) is a consequence of the MaxHom (used by HSSP) favoring local alignments while ClustalW (used by MSAr) favors global alignment. This fact in turn offsets the relative entropy toward the higher values. Nevertheless, the relationship of the relative entropy values remains significantly constant among residues of this particular sequence, allowing for the identification of relevant sites and yielding very similar results, independent of the method used.

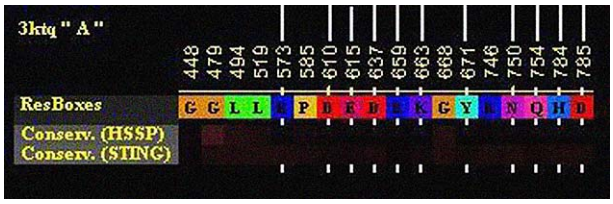


Figure 4. Subset of conserved residues selected by using relative entropy values based on SH_2Q^S alignments and the following [Java]Protein Dossier: relative entropy equal to or less than 8 and reliability higher than 60%. (The reliability is a measure of how many residues are present at a given position in the alignment, effectively showing the ratio of the existing residues occupying such position against the total number of sequences aligned). The residues indicated by solid blank lines correspond to the set selected by using relative entropy values based on the HSSP alignments and the following [Java]Protein Dossier select parameter: relative entropy equal to or less than zero and reliability higher than 60%. Note that by using the SH_2Q^S alignments, seven more residues were detected (G_448, G_479, L_494, L_519, P_585, G_668, and R_746). Also, note that the dark (red) areas below the residues on the lines Conserv. (HSSP) and Conserv. (STING) indicate a very low relative entropy (as opposed to a bright red color used to designate the "red hot" spots of high relative entropy).

Case study 2: HSSP/SH_2Q^S comparative analysis of conservation for residues of angiotensin I-converting enzyme

In this example, we show how usable myMSAr is when a non-PDB structure is analyzed and the degree of conservation of its residues needs to be evaluated. Angiotensin I-converting enzyme (ACE) plays an important role in the cardiovascular homeostasis and regulation of blood pressure, by generating potent vasoconstrictor angiotensin II after cleavage of the C-terminal di-peptide from angiotensin I. In addition, ACE inactivates bradykinin, blocking its hypotensive effects. This enzyme is mainly expressed as somatic and germinal isoforms. The somatic isoform (sACE) has two highly similar active sites, one in the C-domain (sACEc) and another in the N-domain (sACEn), each one having independent catalytic activity. Also, distinct peptide hormone substrates have been found for different ACE enzymes, despite the fact that they show very similar sequences, which can be relevant for the development of selective inhibitors (Fernandez et al., 2003).

Fernandez et al. (2003) have built a molecular model for the N-domain of human sACE (sACEn) in complex with lisinopril. Figure 5 presents the output of the [Java]Protein Dossier (Neshich et al., 2004) displaying only the STING conservation parameter - relative entropy is calculated by using myMSAr, the same process used to generate SH_2Q^S - for the sACEn molecular model. Being a molecular model, there is no entry for this protein in HSSP, but as presented here, the

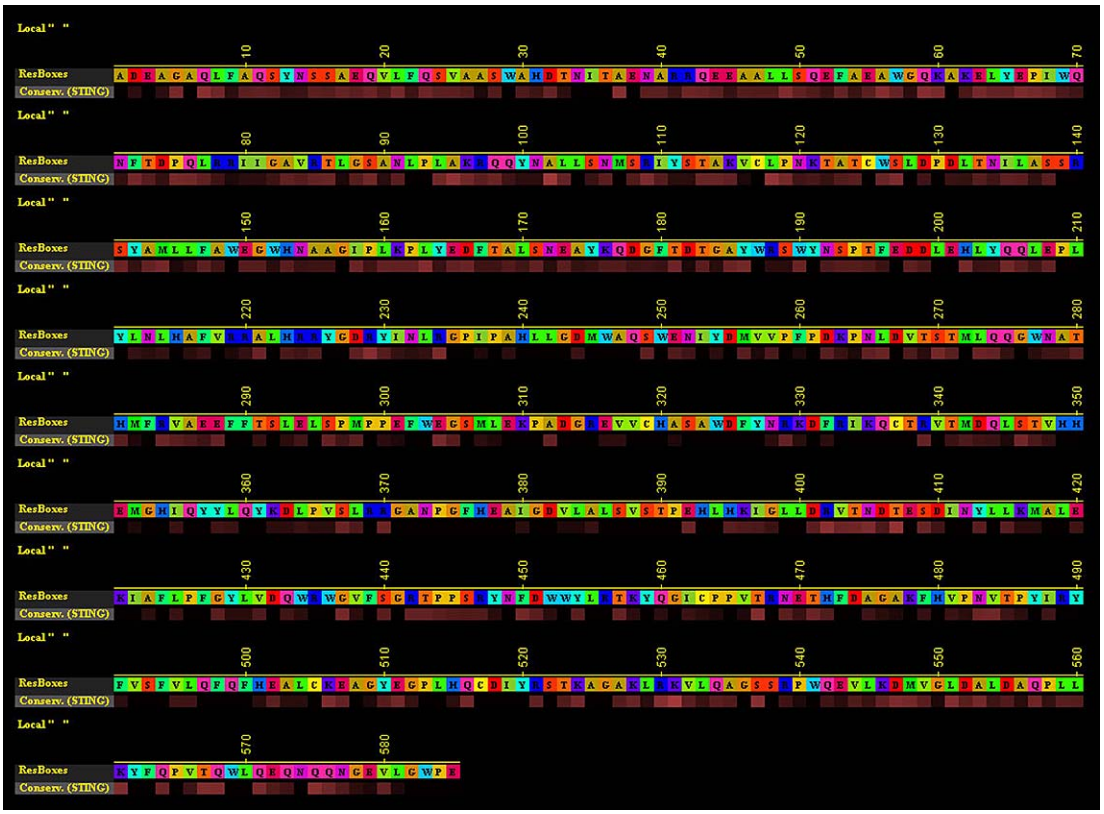


Figure 5. STING conservation for human angiotensin I-converting enzyme N-domain. Conservation is presented as a color-coded shading from dark red - low values of relative entropy - to bright red - high values of relative entropy. The values of relative entropy were calculated by using myMSAr.

conservation parameter calculated by our method may be analyzed in a way similar to those reported by HSSP.

DISCUSSION AND FUTURE WORK

We have presented a new process for building MSAs having a flavor of HSSP. The process is very simple, was implemented by using Perl language and is based on the well-known software packages Blast and ClustalW. Also, a new database - SH_2Q^S has been created by running this process over the entire PDB.

For validation, a dataset of 33,204 query sequences and corresponding HSSP and SH_2Q^S alignments were compared using their degrees of residue conservation. The results indicate that corresponding alignments have equivalent profiles of degree of residue conservation and that this analysis may be performed by using alignments from either HSSP or SH_2Q^S. We also presented two case studies for validation of the results and usefulness of the process: DNA polymerase I and the modeled structure of the ACE N-domain.

Currently, we are developing a new version of myMSAr and the corresponding database - SH_2Q^S based on the Smith-Waterman algorithm (Smith and Waterman, 1981), as described in the original work of Sander and Schneider (1991). In addition, the threshold function

T(L) will be re-adjusted in a way similar to that reported by Rost (1999) but using BLOSUM as the similarity matrix. Finally, we also intend to re-calculate the optimal parameter values used by Blast and Smith-Waterman algorithm to get even closer in terms of approaching the values reported by HSSP.

ACKNOWLEDGMENTS

We would like to thank FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo, Project #1945/01, CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Project #401695/2003-4 and FINEP, Financiadora de Estudos e Projetos, Project #01/08895-0 for supporting this work. We also thank Dr. Jorge Hernandez Fernandez for providing data about conservation for the modeled structure of ACEn.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25: 3389-3402.

Apweiler R, Bairoch A, Wu CH, Barker WC, et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 32: D115-D119.

Beese LS, Derbyshire V and Steitz TA (1993). Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science* 260: 352-355.

Berman HM, Westbrook J, Feng Z, Gilliland G, et al. (2000). The Protein Data Bank. *Nucleic Acids Res*. 28: 235-242.

Cover TM and Thomas JA (1991). Elements of information theory. Wiley Liss Inc., New York, NY, USA.

Fernandez JH, Hayashi MA, Camargo AC and Neshich G (2003). Structural basis of the lisinopril-binding specificity in N- and C-domains of human somatic ACE. *Biochem*. *Biophys*. *Res*. *Commun*. 308: 219-226.

Neshich G, Togawa RC, Mancini AL, Kuser PR, et al. (2003). STING Millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res*. 31: 3386-3392.

Neshich G, Rocchia W, Mancini AL, Yamagishi ME, et al. (2004). JavaProtein Dossier: a novel web-based data visualization tool for comprehensive analysis of protein structure. *Nucleic Acids Res.* 32: W595-W601.

Neshich G, Borro LC, Higa RH, Kuser PR, et al. (2005a). The Diamond STING server. *Nucleic Acids Res.* 33 (Web Server issue): W29-W35.

Neshich G, Mancini AL, Yamagishi ME, Kuser PR, et al. (2005b). STING Report: convenient web-based application for graphic and tabular presentations of protein sequence, structure and function descriptors from the STING database. *Nucleic Acids Res.* 33 (Database issue): D269-D274.

Patel PH and Loeb LA (2000). DNA polymerase active site is highly mutable: evolutionary consequences. *Proc*. *Natl*. *Acad*. *Sci*. *USA* 97: 5095-5100.

Pristovsek P, Lucke C, Reincke B, Ludwig B, et al. (2000). Solution structure of the functional domain of *Paracoccus denitrificans* cytochrome c552 in the reduced state. *Eur*. *J*. *Biochem*. 267: 4205-4212.

Rost B (1999). Twilight zone of protein sequence alignments. *Protein Eng*. 12: 85-94.

Sander C and Schneider R (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet*. 9: 56-68.

Smith TF and Waterman MS (1981). Identification of common molecular subsequences. *J*. *Mol*. *Biol*. 147: 195-197.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22: 4673-4680.

Wall L, Christiansen T and Schwartz RL (1996). Programming Perl. O' Reilly Media Inc., USA.