Protein Structure Alignment through a Contact Topology Profile using SABERTOOTH

F. Teichert[1], U. Bastolla[2], and M. Porto[1]

(1) Institut für Festkörperphysik, Technische Universität Darmstadt,
Hochschulstr. 6-8, 64289 Darmstadt, Germany
(2) Centro de Biología Molecular "Severo Ochoa", (CSIC-UAM),
Cantoblanco, 28049 Madrid, Spain

Abstract: The contact vector (CV) of a protein structure is one of the simplest and most condensed descriptions of protein structure available. It lists the number of contacts each amino acid has with the surrounding structure and has frequently been used e.g. to derive approximative folding energies in protein folding analysis.

The CV, however, is a lossy structure representation, as it does not contain sufficient information to allow for the reconstruction of the full protein structure it was derived from. The loss of information leads to a degeneracy in the sense that a single contact vector is compatible with many different contact matrices, but it has been shown that this degeneracy is nearly fully compensated by the physical constraints protein structure is subject to.

We recently developed the alignment framework 'SABERTOOTH' that is able to generically align connectivity related vectorial structure profiles to compute protein alignments. Here we show that also the CV allows for state-of-the-art alignment quality, just like the elaborated 'Effective Connectivity' profile (EC) that SABERTOOTH currently uses. This simplification leeds to a very simple and elegant approach to structure alignment, which accelerates and generalizes the algorithm we previously proposed.

Furthermore, we conclude from our work that the CV in itself is a useful structure description if its collective properties are called for.

1 Introduction

Alignment of proteins is an every-day remit in many bioinformatics applications and many algorithms exist today that use specialized descriptions of protein structure to solve the problem in a fast and accurate way.

The task, nevertheless, has not been fully solved yet and some improvements are demanded to enhance analyses. Today three different programs are needed for the three different flavours of protein alignment, namely: structural alignment, sequence alignment, and sequence to structure alignment, often referred to as 'threading'. Tailor-made algorithms are available that are specialized for one of these tasks each. Usually, these tools are encumbered with their own often complicated description of protein structure or sequence, respectively. For a user of a software that may result in unforeseeable characteristics and capabilities of the programs, which gets even worse when a combination of two or three different tools are used in the same project.

A desirable alignment tool would comprise all three kinds of alignments using one single algorithm on converging descriptions of protein structure and sequence that should be straightforward in definition and fast to compute.

As a first step into that direction we recently developed the 'SABERTOOTH' alignment framework [TBP07] that allows for the alignment of connectivity related structural profiles. The resulting profile alignment is highly generic and, hence, allows to input different structural and also sequence derived profiles. In a refinement step, actual coordinate data can be used to improve the alignment, if this information is available.

For the profile alignment we relied on the well understood 'Effective Connectivity' (EC) profile [BOPT08] that constitutes a generalization of the Principal Eigenvector of the contact matrix (PE) but allows for the description of complex multi-domain structures, while it is known that the PE nearly exhausively encodes the structural information of small globular folds to the extent contained in the contact matrix [PBRV04]. Besides of the inherent properties that make the EC favourable to other profiles, it is time consuming to compute since diagonalization of the underlying contact matrix is needed.

Here we assess the capacities of the contact vector (CV) of protein structure in our alignment framework. The CV can be understood as an approximation of the EC (see Fig. 1) that is very easy and fast to compute by listing the numbers of contacts each amino acids has with the structure surrounding it. In fact, the CV has a correlation coefficient of $r(\mathrm{EC}, \mathrm{CV}) = 0.94$ with the EC (for EC and CV based on a heavy-atoms contact matrix with distance cut-off $d_{\mathrm{th}} = 4.5\text{Å}$). A potential disadvantage of the CV is that it

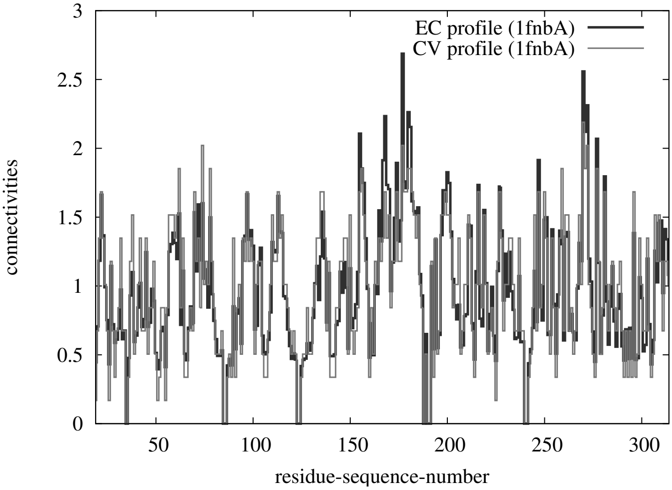

Figure 1: EC and CV profile of the structure with PDB-id '1fnbA' is shown based on a heavy-atoms contact matrix with $d_{\mathrm{th}} = 4.5\text{Å}$. The intriguing correlation between the profiles is obvious.

suffers from degeneracy. If the structure described by a contact matrix is relatively ordered and shows spatial symmetries many contact matrices comply with one and the same contact vector. In [KKVD02] the authors showed that the problem of degeneracy is partly compensated by the distinct properties of native protein structures, i.e. constraints on the protein's backbone like volume exclusion and chemical propensities. Nevertheless, for our application that means that we have to verify the alignment results not only for accuracy in the alignment of related structures but also for the ability to descriminate true and false positives in a mixed set of related and unrelated structures.

Alongside with the move from the EC to the CV, we introduce a second simplification in the profile by changing from a heavy-atoms based contact matrix to one that is derived from the C_α trace of the protein structure, only. In our tests we found very similar performances of the different profiles independent from the choice of coordinates. The C_α description is favourable especially for cases in which the full coordinate information is not available.

To perform the verification of the alignment routine, we firstly show that the alignment results over a test set of related structures are of comparable quality for CV based and the formerly studied EC based alignments. Analysis of a set of alignments of unrelated structures demonstrates the length dependent statistical behaviour giving insight in possible problems with degeneracy.

As a final test we compare the capacities of the different alignments to sort structures according to the 'Structural Classification of Proteins' (SCOP) [MBHC95].

2 Methods

2.1 Contact Matrix, Effective Connectivity Profile, and Contact Vector

The contact matrix of protein structure is a binary symmetric $(N \times N)$-matrix where N equals the number of amino acids in the protein chain. Two amino acids i and j are assigned *in contact* $C_{ij} = 1$ if their spatial distance lies below a threshold d_{th}, or assigned *not in contact* $C_{ij} = 0$ if their distance exceeds the threshold or contacts would be trivial due to the fact that i and j are close along the protein sequence.

The notion of *distance* between amino acids can be defined in many different ways. For the use of structural analyses pairwise distances of the C_α-atoms of the protein's backbone are commonly used, while for problems that depend more on the energetics of side-chain atoms, the minimum of pairwise heavy-atom distances (i.e. other than hydrogen) are preferred.

In this publication we apply both definitions, the EC is based on heavy-atom distances with a d_{th} of 4.5Å whereas we compute the CV from a C_α contact matrix with $d_{\mathrm{th}} = 11$Å, both with three suppressed trivial diagonals, i.e. $C_{ij} = 0$ when $|i - j| < 3$.

Note that this selection is by no means necessary for our analyses, we found that the EC based on C_α distances and the CV based on heavy-atom distances perform nearly as well (data not shown) but chose the particular ones used here since they provide slighly better results. The main reason that C_α atoms are preferable from a practical point of view is that in some applications only the protein's backbone might be known. Furthermore, moving from the truly real-valued/heavy-atoms based EC profile to the integer-valued/C_α based CV accounts for the robustness of our alignment framework.

The contact vector's components CV_i are simply defined as the sum of all elements in row (or column) i of the contact matrix,

$$\mathrm{CV}_i = \sum_{j=1}^{N} C_{ij} \; .$$

The profile actually used in the alignment framework is normalized by dividing its components by the mean value of all connected (i.e. all non-zero) sites to make the components independent of chain length.
The EC, as we defined it in [BOPT08], is a member of the 'Generalized Effective Connectivity' (GEC) family of protein sequence and structure profiles. Like all members of this family, it shares the properties that (a) it maximizes the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$, (b) its mean value is fixed to $\langle c \rangle = 1$ to choose a normalization of its components, (c) its mean square component is fixed to $\langle c^2 \rangle = B > 1$. The corresponding B for the EC is set to $B = \langle \mathrm{CV}^2 \rangle / \langle \mathrm{CV} \rangle^2$ with the contact vector CV_i.
The EC profile can as well be expressed as a weighted sum of eigenvectors of the contact matrix C_{ij}, with weights gradually introducing contributions from more vectors from C_{ij}'s eigensystem when structures described get more modular. Consequently, the values of the components of the EC measure the importance of amino acid i for the global connectivity of the protein structure.
We also showed [BOPT08, TP06] that the EC is nearly identical to the Principal Eigenvector of the contact matrix (which is a member of the GEC family itself), for small single-domain structures with low internal modularity. The PE, in turn, allows for the reconstruction of its contact matrix, hence, its structure with an accuracy comparable to typical X-ray experiments making it a representation of protein structure that is equivalant to atomic coordinates [PBRV04].

2.2 The Alignment Framework 'SABERTOOTH'

The alignment framework introduced in [TBP07] translates the task of finding a proper alignment of two protein structures into the recognition of similar connectivity patterns in the vectorial profiles corresponding to the structures. This analogy is grounded on the observation that the structural profile is conserved in protein evolution, like the overall topology of the protein structure that it describes.
In this way, we can use fast and simple comparison algorithms on the condensed profiles, while relevant non-local properties of protein structure are retained. Moreover, the resulting alignment is little dependent on spurious local similarities that could obliterate the recognition of far homologs. However, these local structural details can be reintroduced in a second step, in order to obtain a more precise structural match.
Following this idea, we developed a structural alignment routine that consists of two steps. First, the alignment of the structural profiles is used to recognize global similarities. Second, a refinement step employs the atomic coordinates in order to improve the local structural superimposition.

2.2.1 Alignment Algorithm

The profile alignment was designed similarly to 'traditional' sequence alignment routines like e.g. dot-matrix alignments. We represented every possible alignment of two proteins by a path through an alignment matrix. Possible alignments were defined as the line-up of two amino acid chains, together with an arbitrary number of inserted gaps of arbitrary length.

The optimum alignment path minimizes a cost function based on the profiles' components and a set of parameters that are analogous to traditional 'substitution probabilities' for alignments and 'open/extend' penalties for gaps. However, in contrast to those, the penalties used here are directly dependent on the structures through their explicit dependence on the profile components.

In order to assess the quality of the resulting alignment, we apply the standard MaxSub routine [SERF00] to the set of aligned residue pairs and compute the optimal rigid body rotation and translation that maximize the spatial superimposition of the two proteins. This allows for the calculation of standard similarity scores based on coordinates and for producing spatial views of the alignment.

Through the MaxSub routine and the set of aligned residues, we derive the optimally superimposed set of coordinates, and from that we compute pairwise distances of all combinations of amino acids connecting the two protein chains. This detailed local information can then be exploited in a second alignment step in order to refine the alignment itself, similar in principle to what other structural alignment algorithms do.

The set of amino acids effectively close in space is analyzed and subsequently used to restrict the possible paths through the alignment matrix, so that the second run searches for the optimal alignment only around these identified groups of close pairs. It incorporates close pair groups into the alignment where unambiguously assigned, it picks out the best choice in cases where more than one alternative is present, and it simply minimizes the path cost as before in areas that are not constrained. Obviously, this kind of refinement is only able to improve the input alignment if the initial spatial superposition was already close to optimal.

After the refinement step, a second run of the MaxSub algorithm is used to obtain the optimal spatial superimposition through which we assess quality and significance of the final alignment. Among other scores a Z-score measuring the statistical significance of the alignment is computed from the Percentage of Structural Identity (PSI) by eliminating the inherent length dependency of the latter.

For more details on the alignment algorithm, cost functions, and parameters please refer to [TBP07].

2.3 Alignment Quality Assessment

In [TBP07] we presented an automatic routine to assess the quality of the alignments produced by our algorithm, as well as of alternative ones produced by well established programs. To do so, we measure the quality of alignments by applying SABERTOOTH and reference tools to a test set of 3566 alignments of distantly related protein pairs by

means of different scores including PSI, contact overlap, and sequence similarity. The structures in the test set are derived from the '29SCOPsf' set described in more detail in [LMLRL$^+$05]. The set consists of 525 structures from 29 SCOP [MBHC95] superfamilies (release 1.69) that constitute a representative collection of common structural motifs. All superfamilies are from different folds of the SCOP classification, and cover the four major SCOP classes all alpha, all beta, alpha+beta, and alpha/beta.
In [TBP07] we could show that SABERTOOTH performs state-of-the-art alignments using the heavy-atoms based EC profile.
In this publication we adopt the same alignment quality assessment routine which makes the results presented here directly comparable to those in our previous publication.

3 Results

3.1 Comparison of Alignment Qualities

The alignment results over the test set of distantly related structures are very similar for EC and CV based alignments. The PSI distributions are depicted in the histograms in Fig. 2 along with the differences in PSI for direct comparison.
The EC profile achieves $\langle \mathrm{PSI}_{\mathrm{EC}} \rangle = 68.2$ while the CV based alignment performes slightly better, resulting in $\langle \mathrm{PSI}_{\mathrm{CV}} \rangle = 69.1$.

3.2 Classification Capacities Assessment

Measuring the capacities of an alignment program to reproduce the SCOP classification constitues a challenging benchmark. Accurately computed alignments are the basis for the assignment of a Z-score that assesses the statistical significance of a given alignment independent of chain lengths. This is only possible if alignments of related structures can be clearly distinguished from unrelated ones.
This attribute can be visualized by an algorithm's behaviour when aligning a set of unrelated structures. The resulting PSI of unrelated pairs plotted versus length of the shorter chain should follow a power-law decay for increasing chain lengths. Figure 3 shows that both profiles perform well in this task and, hence, allow for the definition of proper Z-scores. By fitting a power-law for mean PSI and standard deviation we define the Z-score

$$Z = \frac{\mathrm{PSI} - \langle \mathrm{PSI} \rangle}{\sigma_{\mathrm{PSI}}}$$

with

$$\langle \mathrm{PSI}_{\mathrm{EC}} \rangle = 501.9 \cdot \min\left(N_1, N_2\right)^{-0.714} \text{ and } \sigma_{\mathrm{PSI}_{\mathrm{EC}}} = 541.4 \cdot \min\left(N_1, N_2\right)^{-0.945}$$

$$\langle \mathrm{PSI}_{\mathrm{CV}} \rangle = 493.0 \cdot \min\left(N_1, N_2\right)^{-0.711} \text{ and } \sigma_{\mathrm{PSI}_{\mathrm{CV}}} = 555.6 \cdot \min\left(N_1, N_2\right)^{-0.947} .$$

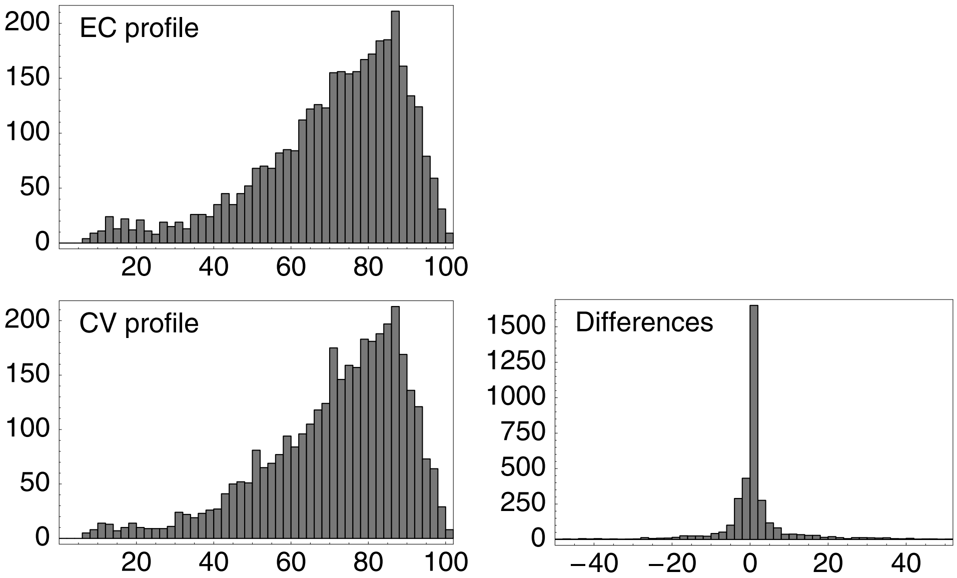

Figure 2: The upper left histogram shows the distribution of PSI values found with EC based alignments as output by SABERTOOTH. The lower left histogram shows the results of the CV based alignments on the same set. The right histogram shows the distribution of the differences $\mathrm{PSI}_{\mathrm{CV}} - \mathrm{PSI}_{\mathrm{EC}}$.

The acutal fold classification capacities are shown in the ROC-plot in Fig. 4. The curve unveils the sensitivity and the generality properties of the Z-score to judge whether the structures in an alignment belong to the same fold in SCOP. The better the classification the larger the area under the curve, i.e. the farther the curve separates from the diagonal line of random guessing.

The set consists of 498 structures that were randomly selected from the 97 largest folds in SCOP (version 1.73) having less than 40% sequence identity. It was assembled by selecting $1/11$ of the structures of all folds with 22 or more members in the ASTRAL40 [CHW$^+$04] database. All-vs-all alignment generates 123753 alignments of protein chains with known SCOP relation.

4 Conclusions

We could show that the very condensed and simple but also lossy representation of protein structure as a contact vector still contains sufficient information to perform structural alignments. Furthermore, the behaviour with unrelated structures is very similar to that of the more sophisticated EC profile. This means that the degeneracy the CV suffers from does not play a major role for this application. This remains true even after reduction of the input data from heavy-atom coordinates to a C_α description.

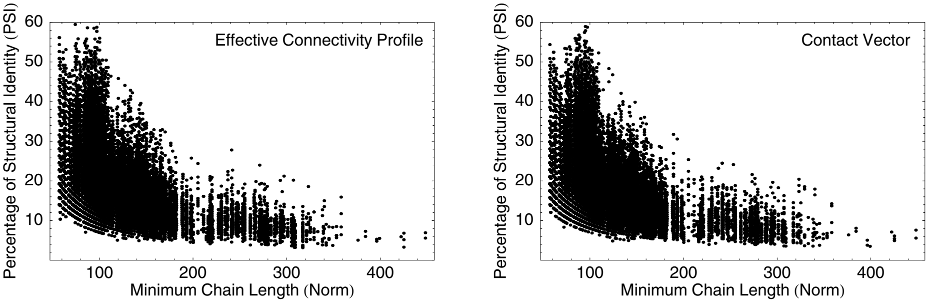

Figure 3: PSI versus minimum chain length for EC (left figure) and CV (right figure). Both plots show the same power-law length dependence for applying SABERTOOTH on a set of unreated structures. Well defined Z-scores can be computed for both distributions.

The slightly superior performance of the CV in our alignment framework, in comparison with the EC, together with its lower computational cost persuaded us to move to the CV as the standard structure representation for our alignment program SABERTOOTH (refer to `http://www.fkp.tu-darmstadt.de/sabertooth/`).
Moreover, from our analyses we conclude that the CV, just as being so simple to compute, might be a better description for analyzing collective properties of protein network topology than one could expect.

5 Funding

FT and MP gratefully acknowledge generous financial support from the Deutsche Forschungsgemeinschaft via project PO 1025/1 and from the Deutscher Akademischer Austauschdienst via project D/06/12858. UB acknowledges financial support from the Spanish Education and Science Ministry through the Ramón y Cajal program and the grant no. BIO2005-05786 and from CSIC through the Acciones Integradas program.

6 Acknowledgements

We are publishing the present paper in remembrance of Angel Ramirez Ortíz (1966-2008), with whom we had the privilege to have inspiring discussions and numerous hints that we gratefully acknowledge, while we deeply miss his friendship and advice.
We thankfully acknowledge Alejandra Leo-Macías for the manually refined 29SCOPsf set.

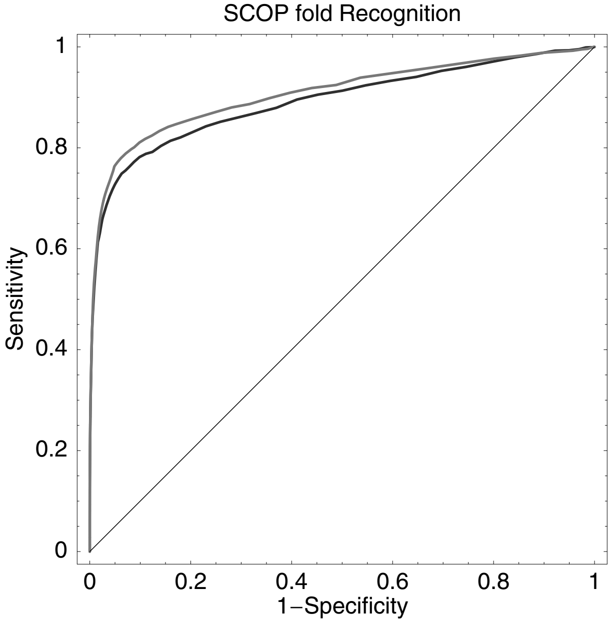

Figure 4: Fold recognition capacities for SABERTOOTH alignments using the EC profile (blue curve) and the CV profile (red curve). It turns out the the CV performs slightly better here than the EC.

References

[BOPT08] Ugo Bastolla, Angel R. Ortíz, Markus Porto, and Florian Teichert. Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences. *PROTEINS: Structure, Function, and Bioinformatics*, 2008. (in print; doi: 10.1002/prot.22113).

[CHW+04] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner. The ASTRAL compendium in 2004. *Nucleic Acids Research*, 32:189–192, 2004.

[KKVD02] A. Kabakçioglu, I. Kanter, M. Vendruscolo, and E. Domany. Statistical properties of contact vectors. *Physical Review E*, 65(4):41904, 2002.

[LMLRL+05] Alexandra Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, and Angel R. Ortiz. An analysis of core deformations in protein superfamilies. *Biophys J*, 88(2):1291–1299, 2005.

[MBHC95] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[PBRV04] Markus Porto, Ugo Bastolla, H.-Eduardo Roman, and Michele Vendruscolo. Reconstruction of Protein Structures from a Vectorial Representation. *Physical Review Letters*, 92(21):218101, 2004.

[SERF00] Naomi Siew, Arne Elofsson, Leszek Rychlewski, and Daniel Fischer. MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, 2000.

[TBP07] Florian Teichert, Ugo Bastolla, and Markus and Porto. SABERTOOTH: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, 8:425, 11 2007. Available online at `http://www.fkp.tu-darmstadt.de/sabertooth/`.

[TP06] F. Teichert and M. Porto. Vectorial representation of single-and multi-domain protein folds. *The European Physical Journal B-Condensed Matter and Complex Systems*, 54(1):131–136, 2006.