ELECTRONIC SUPPLEMENTARY MATERIAL (ESM)

# Novel protein folds and their non-sequential structural analogues

Aysam Guerler (guerler@chemie.fu-berlin.de) and Ernst-Walter Knapp* (knapp@chemie.fu-berlin.de)
Department of Chemistry and Biochemistry, Freie Universität Berlin, Fabeckstrasse 36a, 14195, Berlin, Germany

## Algorithmic details of GANGSTA+

GANGSTA+ as well as GANGSTA (Kolbeck et al. 2006) have in common to align protein structures hierarchically starting on the secondary structure level (first stage) before transforming the results to residue level. In contrast to the former GANGSTA, which used a genetic algorithm, GANGSTA+ uses a combinatorial approach on the secondary structure level to evaluate similarities between two protein structures based on contact maps. For this purpose we consider the secondary structure elements (SSE), i.e. α-helices and β-strands (ignoring loops and coils), of the two proteins to be compared and make pairwise assignments of SSEs (belonging to the same type) from the two structures according to a specific similarity measure, based on the GANGSTA objective function (GOF) (Kolbeck et al. 2006). In GANGSTA+ the bookkeeping of these SSE assignments is performed in a linear array, **map**, whose indices **i** enumerate the SSEs of the polypeptide chain (A) under consideration (source protein, which should be the smaller of the two proteins to facilitate the computation), while the integer values $X_i$ in the array label the SSEs of the polypeptide chain (B) (target protein, which generally is the larger of the two proteins) taken from a database of protein structures. To allow for gaps in chain A, the $X_i$ in the SSE map can also adopt the value **G** denoting that the SSE **i** in chain A is not assigned to a SSE in chain B.

Different SSE assignment modes can be used with GANGSTA+. The assignment of SSEs can be performed respecting the sequential order of the SSEs in the polypeptide chains of the considered protein pair (sequential alignment) or ignoring this order (non-sequential alignment). Furthermore, SSE pairs can optionally be aligned in reverse orientation.

The combinatorial approach of GANGSTA+ aims to find the SSE **map**, which maximizes the GOF score (Kolbeck et al. 2006). Therefore, we initially construct a list (**maplist**), of $\mathbf{M_{comb}}$ members containing all possible incomplete SSE **maps** involving two SSE pairs only, which we call **2-tuple** SSE **maps**. These **2-tuple** SSE **maps** contain only pairs of SSEs belonging to the same type, i.e. α-helix or β-strand. Ignoring the difference between α-helical and β-strand SSEs, which overestimates the total number of possible **2-tuple** SSE maps, $\mathbf{M_{comb}}$ would for instance be

$$M_{comb} = \tfrac{1}{2}[n_A(n_A - 1)\ n_B(n_B - 1)], \quad (1)$$

where $\mathbf{n_A}$ and $\mathbf{n_B}$ are the number of SSEs in chain A and B, respectively. We sort all **2-tuple maps** in the **maplist** according to the GOF score and consider only the $\mathbf{N_{maplist}} = \min(\mathbf{M_{comb}} * \mathbf{R_{ratio}}, \mathbf{N_{max}})$ (default values are $\mathbf{R_{ratio}} = 0.5$ and $\mathbf{N_{max}} = 1000$) highest ranked maps according to the GOF score. In the subsequent procedure the size of the ordered **maplist** is limited to $\mathbf{N_{maplist}}$. In an iteration procedure, higher order **n-tuple maps** are generated by merging two maps $\mathbf{map_k}$ and $\mathbf{map_l}$ from the ordered **maplist** starting from the top of the list with the highest ranked tuple maps. In the outer loop **k** runs from 1 to $\mathbf{N_{maplist}}$, while in the nested inner loop **l** runs from **k+1** to $\mathbf{N_{maplist}}$. SSE **maps** with conflicting SSE pair assignments are not merged, but skipped. The resulting new map is considered to be successful, if it possesses more assigned SSE pairs, a better GOF score than each of the two original maps and has not been generated before. The latter is checked by using a search tree generated in parallel with the algorithm. If $\mathbf{N_{maplist}}$ successful merged maps are generated and placed in a second intermediate storage list the iteration cycle terminates. While the inner loop index runs over the full **maplist**, the outer loop index **k** can reach also large values before the intermediate maplist is filled, since attempts to merge two maps are often not successful. Now the filled intermediate list is merged with the **maplist**. The resulting $2*\mathbf{N_{maplist}}$ maps are ordered and the top $\mathbf{N_{maplist}}$ ranked maps are placed in the updated **maplist** and a next iteration cycle can start. Up to three iteration cycles are performed, which allow to generate **maps** assigning up to 16 SSE pairs. For medium size proteins this often results in a complete SSE assignment. For larger proteins the SSE assignment is com-

pleted in the third step of structure refinement as explained below. However, the iteration may be terminated earlier, if it was not possible to generate a new successful map in an iteration cycle, which is often the case for small proteins. Finally the ordered **maplist** contains in the first positions the best structure alignment results in terms of assigned SSE pairs, which can be used to perform structure alignment on the residue level in the second step as described below.

The residue level alignment (second stage) follows the secondary structure level optimization and is applied on the $\mathbf{N_{map}}$ highest ranked SSE **maps** (default value is $\mathbf{N_{map}} = 50$) of the pair of protein structures to be aligned. To obtain an initial common set of atomic coordinates for both proteins, we define pairwise attractive interactions of the $C_\alpha$ atom pairs in terms of inverse Lorentzians $V_{Lorentz}(\vec{r}_i - \vec{r}_j)$ describing the interactions between atoms **i** and **j**

$$V_{Lorentz}(\vec{r}) = \left[\vec{r}^{\,2} + 0.01\text{Å}^2\right]^{-1}. \quad (2)$$

These interactions apply only for $C_\alpha$ atoms that belong to equivalent SSE pairs but to different proteins. With this artificial energy function, which describes the attraction of equivalent SSEs of the two aligned proteins, we perform an energy minimization similarly as it was done recently in an application of protein-ligand docking (Guerler et al. 2007). Subsequently, all $C_\alpha$ atoms of the aligned protein pair are projected on lattice points of the same 3D grid with 1.0 Å resolution keeping track of the protein and the SSE id to which they belong. Note that for this procedure the $C_\alpha$ atoms in all SSEs of the two proteins are considered. Two $C_\alpha$ atoms (i, j) belonging to different proteins are in contact, if their Chebyshev distance $D_{Cheb}$

$$D_{Cheb}(i, j) = \max[|x_i - x_j|, |y_i - y_j|, |z_i - z_j|] \quad (3)$$

of the assigned grid points is less than or equal to $D_{Cheb}(max)$. In this grid representation of the structure alignment the number of $C_\alpha$ atom pair contacts between SSEs of equivalent type are counted. According to the number of $C_\alpha$ atom pair contacts the SSE pair assignment is repeated (third stage), resulting in a new SSE **map**, which for large proteins often results in an enlarged set of assigned SSE pairs and may also lead to reassignments of SSEs in some cases. Thus, possible incomplete SSE assignments from the first stage optimization are now completed. Based on this new map the energy minimization of the $C_\alpha$ atom pairs between equivalent SSEs from different proteins is repeated using the inverse Lorentzians, eq. (2). Now, pairs of equivalent residues in the revised protein structure alignment are assigned in a correlated way that extends the assignment of residues even beyond the boundaries of SSEs extending in the loop and coil regime, if the Euclidian distances of the corresponding $C_\alpha$ atom pairs are smaller than a cut-off distance $D_{Euler}(max)$. The two cut-off distances $D_{Euler}(max)$ and $D_{Cheb}(max)$ are related and should be chosen such that $D_{Euler}(max) > D_{Cheb}(max)$. The cut-off distances limit the maximum deviation that equivalent $C_\alpha$ atom pairs can have in a pair of aligned protein structures and limit also the RMSD of all aligned $C_\alpha$ atom pairs, which as a consequence is much below this maximum value $D_{Euler}(max)$. Hence, small $D_{Euler}(max)$ and $D_{Cheb}(max)$ lead to high quality protein structure alignments, which at the same time can involve only a small number of aligned residues. In the present application, we are interested in alignments of relatively high quality and use therefore $D_{Cheb}(max) = 2.0$ Å and $D_{Euler} = 4.0$ Å leading to RMSDs, which are typically below 2.5 Å (see for an example figure 2C and 3C). Finally, the Kabsch algorithm (Kabsch 1976) is applied to minimize the RMSD of all equivalent $C_\alpha$ atom pairs in the two aligned protein structures. In the present study, we search for sequential and non-sequential structure alignments and allow for pairwise SSE alignments also in reverse orientation. To ensure that protein structure alignments cover a sufficiently large portion of the structures we accept only alignment results involving at least 50% of the SSEs of the smaller of both proteins (the source protein) and more than 40 aligned residues ($N_{aligned} > 40$).

## Non-sequential protein structure alignments with same SSE orientation

Protein structure alignments of novel protein folds and the ASTRAL40 (Murzin et al. 1995) database generated with GANGSTA+. SSE assignments between two protein chains are restricted to be oriented in the same direction. Figure S1A, S1B shows **2JMK** (Koo et al. 2007) aligned on **1VJU** (SGPP) with 45 residues and a RMSD of 2.4 Å. The overall result of the corresponding database scan is depicted in figure S1C. Figure S2A, S2B shows **2ES9** (Benach et al. 2005) aligned on **1H6K** (Mazza et al. 2001) with 48 residues and a RMSD of 2.2 Å. Figure S2C shows the corresponding overall result of the database search with **2ES9** and the ASTRAL40 database.

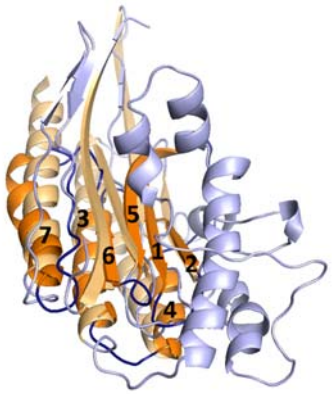


**Figure S1A** Protein structure alignment with GANGSTA+. New fold **2JMK** (Koo et al. 2007) (dark colors, blue for loops and orange for SSEs) aligned on **1VJU** (SGPP) (light colors, blue and orange) yielding the RMSD = 2.4 Å with 45 aligned residues and 7 aligned SSEs. The aligned SSEs of **2JMK** (**1VJU**) are represented in dark (light) orange, not aligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially referring to the reference protein **2JMK**.

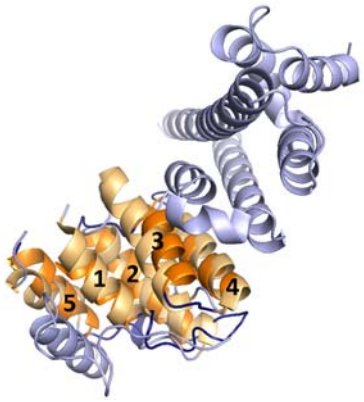


**Figure S2A** Protein structure alignment with GANGSTA+. New fold **2ES9** (Benach et al. 2005) (dark colors, blue for loops and orange for SSEs) aligned on **1H6K** (Mazza et al. 2001) (light colors, blue and orange) yielding the RMSD = 2.2 Å with 48 aligned residues and 5 aligned SSEs. The aligned SSEs of **2ES9** (**1H6K**) are represented in dark (light) orange, not aligned parts (SSEs and loops) are in dark (light) blue. The SSEs are numbered sequentially referring to the reference protein **2ES9**.

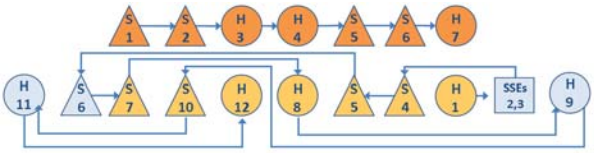


**Figure S1B** Connectivity graph of protein structure alignment. Aligned SSE pairs are on top of each other. SSEs are numbered in sequential order; H: α-helices, circles; S: β-strands, triangles; connecting loops, blue arrows. Top part: **2JMK** (all SSEs in dark orange); bottom part: **1VJU** (aligned SSEs in light orange, not aligned SSEs in light blue). The SSEs are numbered sequentially. SSE pairs assigned in reverse orientation are marked by arrows. All SSE pairs are assigned in the same orientation.

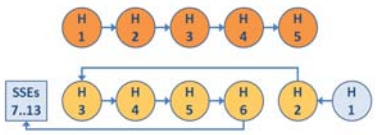


**Figure S2B** Connectivity graph of protein structure alignment. Aligned SSE pairs are on top of each other. SSEs are numbered in sequential order; H: α-helices, circles; S: β-strands, triangles; connecting loops, blue arrows. Top part: **2ES9** (all SSEs in dark orange); bottom part: **1H6K** (aligned SSEs in light orange, not aligned SSEs in light blue). The SSEs are numbered sequentially. SSE pairs assigned in reverse orientation are marked by arrows. All SSE pairs are assigned in the same orientation.

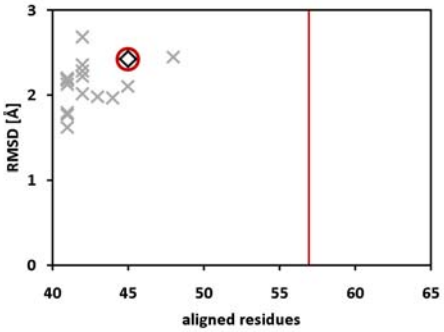


**Figure S1C** Diagram correlating the number of aligned residues with the RMSD for the structure alignment results of GANGSTA+ with respect to **2JMK** and the ASTRAL40 dataset (diamonds mark alignments involving all SSEs of **2JMK**, crosses mark incomplete alignments). All results with more than 40 aligned residues are displayed. The structure of **2JMK** consists in seven α-helices, which comprise a total of 57 residues, marked by the red line. The red circle marks the structure alignment with **1VJU** shown in the figures S1A, S1B.

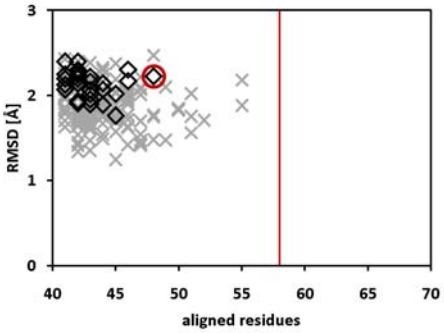


**Figure S2C** Diagram correlating the number of aligned residues with the RMSD for the structure alignment results of GANGSTA+ with respect to **2ES9** and the ASTRAL40 dataset (diamonds mark alignments involving all SSEs of **2ES9**, crosses mark incomplete alignments). All results with more than 40 aligned residues are displayed. The structure of **2ES9** consists in five α-helices, which comprise a total of 58 residues, marked by the red line. The red circle marks the structure alignment with **1H6K** shown in the figures S2A, S2B.

## REFERENCES


Benach, J., Abashidz, E.M., Jayaraman, S., Rong, X., Acton, T.B., Montelione, G.T., and Tong, L. 2005. Crystal structure of Q8ZRJ2 from salmonella typhimurium. NESG TARGET STR65. *to be published*.

Guerler, A., Moll, S., Weber, M., Meyer, H., and Cordes, F. 2007. Selection and flexible optimization of binding modes from conformation ensembles. *Biosystems*.

Kabsch, W. 1976. A solution of the best rotation to relate two sets of vectors. *Acta Crystallographics* **32:** 922.

Kolbeck, B., May, P., Schmidt-Goenner, T., Steinke, T., and Knapp, E.W. 2006. Connectivity independent protein-structure alignment. *BMC Bioinformatics* **7**.

Koo, B.K., Jung, J., Jung, H., Nam, H.W., Kim, Y.S., Yee, A., and Lee, W. 2007. Solution structure of the hypothetical novel-fold protein TA0956 from Thermoplasma acidophilum. *Proteins* **69:** 444-447.

Mazza, C., Ohno, M., Segref, A., Mattaj, I.W., and Cusack, S. 2001. Crystal structure of the human nuclear cap binding complex. *Mol.Cell* **8:** 383-396.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP. *J. Mol. Biol.* **247:** 536-540.

SGPP, S.G.o.P.P.C. Primary Citation Protozoa, Structural Genomics of Pathogenic Coproporphyrinogen III oxidase from Leishmania major. *to be published*.