Fast Protein Fold Recognition via Sequence to Structure Alignment and Contact Capacity Potentials

by

Nickolai N. Alexandrov[1], Ruth Nussinov[2,3] and Ralf M. Zimmer[4]

[1]Laboratory of Mathematical Biology, NCI–FCRF
P.O. Box B, Frederick, MD 21702–1201, USA
Email: nicka@ncifcrf.gov Phone/Fax: +01 301 846 6542 / 5598
[2]Sackler Inst. of Molecular Medicine, Faculty of Medicine
Tel Aviv University, Tel Aviv 69978, Israel
[3]Laboratory of Mathematical Biology, SAIC, NCI–FCRF
Bldg 469, Rm 151, Frederick, MD 21702–1201, USA
[4]Institute for Algorithms and Scientific Computing, GMD–SCAI
Schloß Birlinghoven, P.O. Box 1316, D 53754 Sankt Augustin, Germany
Email: Ralf.Zimmer@gmd.de Phone/Fax: +49 2241 14 2818 / 2656

Abstract

We propose new empirical scoring potentials and associated alignment procedures for optimally aligning protein sequences to protein structures. The method has two main applications: first, the recognition of a plausible fold for a protein sequence of unknown structure out of a database of representative protein structures and, second, the improvement of sequence alignments by using structural information in order to find a better starting point for homology based modelling. The empirical scoring function is derived from an analysis of a non–redundant database of known structures by converting relative frequencies into pseudoenergies using a normalization according to the inverse Boltzmann law. These – so called contact capacity – potentials turn out to be discriminative enough to detect structural folds in the absence of significant sequence similarity and at the same time simple enough to allow for a very fast optimization in an alignment procedure.

1 Introduction and Problem Definition

Predicting protein structure from the protein sequence is one of the most challenging problems of molecular biology with many applications and consequences for theory and experiment. We are interested in the following quite simple instance of this problem: Given a sequence of unknown structure and a database of representative folds, identify the most plausible fold for the sequence if there is one and assess the quality or reliability of the proposed structure.

Towards this goal we applied the following method: Using simple empirical potentials we optimize mappings of residues of the sequence onto structural positions of any of the proposed folds (so called sequence–structure alignments or threadings [2, 5, 18]). The resulting alignments are then evaluated and ranked according to the potential and the statistical significance of the best alignment is estimated in comparison with the other alignments.

In an approach related to [2, 14], we use simple environment dependent potentials – so called contact capacity potentials (CCPs)–, which do not explicitly depend on the actual contact partner. The latter dependency would destroy the so called prefix optimality principle and, thus, prohibit fast optimization via dynamic programming. It is generally believed that the objective function for sequence–structure alignment needs to include pairwise contact energy terms to get high quality alignments and good discrimination between appropriate and inappropriate folds. An experiment demonstrates that the major contributions for native fold recognition in the sippl–test are various types of contact capacity instead of detailed pairwise contact energies. We also show that more involved CCPs do not improve much in the sippl–test as compared to simple CCPs and secondary structure preferences.

From these experiments we conclude that for the even less precise energy estimations employed in threading with gaps it should be possible to restrict oneself to CCPs thus allowing for fast and exact optimization. Additionally, in current procedures, there is some discrepancy in having quite significant fold recognition but only poor alignments. Our goal is to get both good alignments and good discrimation with a fast dynamic programming optimization procedure. Therefore, we adapted the dynamic programming to deal with CCPs and to account for position and secondary structure dependent costs, especially gap costs, and for averaging scores for certain matches over a window centered at this match. We implemented these options with several versions of this kind of potentials and evaluated the effects on fold recognition and detailed alignments using the **ToPLign** procedures [11].

2 Methods

2.1 Potentials

We used statistically derived potentials computed from a non–redundant set of representative protein structures suggested by Bauer&Beyer[1]. This set was obtained from the list of 185 non–homologous structures proposed by Hobohm&Sander [10] by eliminating membrane and virus proteins. As our empirical free energy function we use a sum of three terms: secondary structure preferences, pairwise contact potentials, and contact capacity potentials.

2.1.1 Secondary Structure Preference

For each amino acid we assigned one of three types of secondary structures (SS): alpha, beta and other. Assignment was made based on the similarity of the 5–residue C_α–trace fragment from the structure to the typical α–helix or β–strand. From the total number of 66634 amino acids, 24970 were assigned to class alpha, 17403 to class beta, and 24261 to class other. The secondary structure preference (SSP) of amino acid i to be in secondary structure class s are calculated from the following formula:

$$\mathrm{Pss}(i,s) = -log\frac{N(i,s)}{\langle N(i,s)\rangle}, \qquad \langle N(i,s)\rangle = \frac{N(i)*N(s)}{N},$$

where $N(i,s)$ is the actual number of amino acids i in secondary structure conformation s, $\langle N(i,s)\rangle$ is the expected number of the residue i to be in SS class s, $N(i)$ is the number of the amino acids i, $N(s)$ is the number of amino acids in conformation s, and N is the total number of amino acids.
The potential we obtained is similar to those obtained previously by many authors, e.g. Chou and Fasman [4], and is summarized in table 1.

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALPHA	-33	44	0	-24	-1	38	4	1	-9	-24	-23	4	78	-22	-17	14	23	20	-2	19
BETA	32	-15	27	33	-18	11	0	-39	9	-1	9	27	6	22	15	7	-21	-45	-12	-22
OTHER	22	-20	-15	7	16	-31	-4	37	3	32	22	-19	-44	10	8	-16	-3	23	12	0

Table 1: Secondary structure potentials (multiplied by 100 for clarity)

2.1.2 Pairwise Contact Potentials

This type of potential was initially suggested by Miyazawa and Jernigan [12]. We defined a pair of residues as being in contact, if the distance between C_β–atoms is less than 7.0 Å. The coordinates for a fake C_β–atom for Glycine were calculated from the backbone. The contact potentials for amino acids i and j were calculated as follows:

$$\mathrm{Pc}(i,j) = -log\frac{N(i,j)}{\langle N(i,j)\rangle}, \qquad \langle N(i,j)\rangle = \frac{N(i)*N(j)}{N}$$

where $N(i,j)$ is the actual number of residues i and j in contact, $\langle N(i,j)\rangle$ is the expected number computed from $N(i) = \sum_j N(i,j), N(j) = \sum_i N(i,j)$, and $N = \sum_{i,j} N(i,j)$ These potentials are in good agreement with previously derived contact potentials.

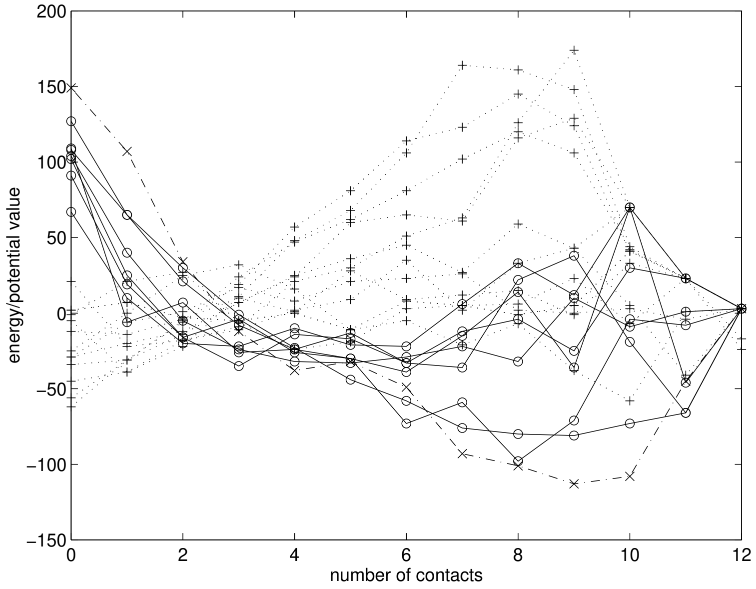

Figure 1: Contact capacity potentials for hydrophobic (circled), polar (dotted), and Cysteine residues (dashed).

2.1.3 Contact Capacity Potentials

This new type of potentials was introduced to account for the hydrophobic contribution to the free energy. Contact capacity characterizes the ability of residues to make a certain number of contacts with any other residues. Obviously, hydrophobic residues should have more contacts than polar residues. For each type i of amino acid, we derived its ability to form k contacts $\mathrm{Pcc}(i,k)$as:

$$\mathrm{Pcc}(i,k) = -log\frac{N(i,k)}{\langle N(i,k)\rangle}, \qquad \langle N(i,k)\rangle = \frac{N(i) * NC(k)}{N}$$

where $N(i,k)$ is the number of residues i having k contacts. The expected number $\langle N(i,k)\rangle$ is calculated using $N(i)$ the number of residues i, $NC(k)$ the number of residues having k contacts, and N the total number of residues.
We used several different variations of the contact capacity potentials:
Long-range and local contact capacity potentials (CCP): We divided the contacts and, correspondingly, the contact capacity potentials into two categories: local and long–range potentials. We consider a contact as local if there are less than five residues in the sequence between the two residues in contact. There is a clear correlation between residue hydrophobicity and long range contact capacity potentials (figure 1). Local contact capacity shows some correlation with secondary structure preferences: obviously, those residues,

which have a preference to be in α–helical conformation tend to have more local contacts.

Secondary structure dependent CCP (SSCCP): The ability of the residues to make contacts may depend on the secondary structure: residues in α–helices have less vacant surrounding space for contacting residues than residues in β–strands. Thus, we have derived and tested secondary structure dependent potentials, having now 6 types of contact capacity potentials: two sequence separations (local and long–range) multiplied by three types of the SS. There is a significant difference between the long–range alpha CCP and the long–range beta CCP (tables 2 and 3).

#contacts:	0	1	2	3	4	5	6	7	8	9
ALA	8	35	39	-10	-40	-36	-22	-3	0	5
CYS	95	78	26	-36	-64	-24	-73	-111	-133	5
ASP	-61	-20	16	72	104	115	79	53	40	5
GLU	-60	-20	22	54	95	104	134	99	40	5
PHE	107	11	-44	-49	-19	20	11	54	1	5
GLY	-14	13	32	52	-3	-35	-59	-85	-31	-47
HIS	31	-43	-46	14	35	77	59	-34	40	5
ILE	108	38	-7	-34	-44	-46	-60	-33	-34	-26
LYS	-49	-38	-8	60	134	165	130	122	40	5
LEU	109	27	-21	-41	-41	-34	-12	-22	22	5
MET	82	2	-18	-34	-13	-42	-18	63	-17	5
ASN	-36	-1	6	21	39	50	40	-13	1	5
PRO	-28	10	2	31	37	19	-37	-14	-96	5
GLN	-25	-25	-10	30	33	65	73	85	40	5
ARG	-6	-44	-23	5	80	96	70	87	40	5
SER	-26	1	18	31	8	-5	17	-5	-14	5
THR	-9	13	7	10	12	-31	-18	-16	-21	5
VAL	106	48	21	-38	-48	-61	-79	-62	9	5
TRP	99	-39	-25	-20	-7	-35	81	122	40	5
TYR	100	-5	-33	-31	-33	0	11	70	-11	5

Table 2: Long–range CCP (*times* 100) for residues in alpha–conformation

Conditional CCP (CCCP): Assuming that the local CCP reflects a type of SS preference in protein structures, we can replace the SS dependent CCP with a conditional CCP. The local CCP in this case do not depend on the SS, whereas the long–range CCP depend on the number of local contacts. An example of this type of long–range contact capacity potentials is shown in table 4. Conditional CCP are consistent with the hierarchical model of protein folding [15], when an initial SS formation is followed by the formation of the long–range contacts.

Distance–dependent CCP (DCCP): Sippl introduced distance–dependent contact potentials, gaining excellent performance in the Sippl–test. We tried to exploit the same idea to improve the results of the Sippl–test with CCP. We have introduced six distance intervals: 4–5 Å, 5–6 Å, 6–7 Å, 7–8 Å, 8–9 Å, and 9–10 Å and 6 sequence separations between contacting residues: 1, 2, 3, 4, 5, and 6 or more. The total number of the parameters for each amino acid with

#contacts:	0	1	2	3	4	5	6	7	8	9	10	11	12
ALA	9	26	23	10	-5	0	-9	-19	-31	-3	-3	9	-24
CYS	197	124	71	8	-1	-4	-18	-51	-67	-77	-47	-48	3
ASP	-59	-59	-38	-4	-6	19	68	104	89	58	47	9	-31
GLU	-55	-41	-58	-28	3	26	92	91	123	103	9	9	3
PHE	91	35	24	7	-4	-19	-33	-21	18	-14	47	9	3
GLY	-74	-21	9	17	33	0	25	4	10	-8	-67	-14	3
HIS	7	-1	13	-43	-21	-6	45	22	27	76	47	9	3
ILE	153	105	55	30	19	1	-57	-40	-73	-39	24	-13	3
LYS	-40	-50	-56	-28	2	9	89	125	119	141	18	9	3
LEU	101	90	55	21	-3	-33	-36	-22	-35	4	-7	-12	3
MET	86	50	12	7	-9	-17	-45	-2	10	66	47	9	3
ASN	-41	-63	-42	0	5	34	48	54	52	18	47	9	3
PRO	-17	-54	-19	-6	-10	43	24	29	39	-3	9	9	3
GLN	-36	-48	-25	-41	-12	50	47	53	114	87	47	9	3
ARG	-24	-9	-5	-45	-27	-2	66	51	108	141	47	9	3
SER	-42	-42	-28	1	-3	36	40	27	17	29	-23	9	3
THR	-24	-6	-26	-23	0	5	42	27	20	16	47	9	3
VAL	81	77	64	33	8	-14	-28	-54	-43	-64	-37	-7	3
TRP	127	65	35	-13	-20	20	-40	-35	10	13	-36	9	3
TYR	103	52	23	41	-10	-35	-41	-10	17	-28	47	9	3

Table 3: Long–range CCP (*times* 100) for residues in beta–conformation

	# of long-range contacts												
	0	1	2	3	4	5	6	7	8	9	10	11	12
# of local contacts													
0	218	135	109	30	-51	-17	-29	-93	-91	-91	-82	-56	
1	219	218	67	20	4	0	-12	-95	-96	-86	-99	3	
2	153	74	4	-18	-24	-48	-78	-78	-94	-116	18	7	1
3	109	105	20	-51	-38	-48	-54	-77	-26	9	1		1
4	110	57	47	3	-57	-27	-54	-55	9	1			
5	17	48	-18	-26	-77	-51	-67	7					
6	37	41	-66	-89	-41	9	5						
7	1	11	5	1		1	1						

Table 4: Long–range conditional CCP for Cysteine (multiplied by 100)

a certain number of contacts equals to 3*6*6=108 (3 types of SS, 6 distance and 6 sequence separations).

Angle–dependent CCP (ACCP): A more detailed version of CCP could be introduced via division of each distance interval into 6 segments, depending on the orientation of the contacting residue (figure 2). The total number of parameters for each amino acid with a certain number of contacts in this case is 3*6*6*6=648 (3 types of SS, 6 distance and 6 sequence separations, 6 angle segments).

2.2 Assessing Potentials and Threading Methods

To assess the performance of our simple potentials and the associated optimization procedures for fold recognition, we employ the following tests proposed in the literature:

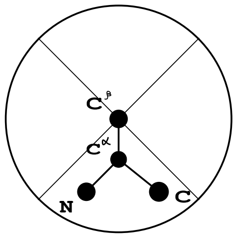

Figure 2: 2–dimensional picture of the contact region, divided into four areas of contacts. The position of the areas of contacts depends on the coordinates of the backbone atoms. In 3D we use 6 areas of contacts.

shuffle–test [3]: The most naive and simple test is a standard statistical test: Given some 'native' score of an optimization procedure, compute the optimal score of many randomized/rotated/permutated inputs, which are required to have the same amino acid composition, and calculate the native score in terms of standard deviations of the resulting randomized scores.
Sippl–test [17]: Sippl proposed and used the following test: Given a sequence S of length n and a database of structures, mount the sequence onto all possible structures of length at least n without gaps, i.e. cutting fragments of length n out of all longer structures and assuming their sequence to be S. Evaluate the potential score for all of these combinations, assume a normal distribution of the scores, and calculate the score of the native combination in terms of standard deviations. For this test keep in mind that only one sequence–structure pair corresponds to the native combination, all the other combinations are simply wrong structures for the sequence or even incomplete protein structures. Sippl claims that this test could be reasonably interpreted: The sequence is a 'real' physical system, which tends to adopt the minimum energy conformation. Therefore, it makes sense to look through conformational space and select those conformations which minimize the free energy of the respective sequence–structure pair.
Threading–test [2, 3, 5]: This test is considered to be the most realistic one: Using the respective method and potential try to align a given sequence as well as possible to any fold of a fold database, evaluate, score and rank the resulting alignments. The native (identity) threading should be the best alignment of the sequence onto its native fold, and the score of this combination should be better than the score of all non–native combinations. Also, similar folds should be ranked quite high in the list and dissimilar ones ranked low.

2.3 Aligning using Contact Capacity Potentials and Sequence Information

2.3.1 Alignments: Modes and Algorithms

For doing threading with contact capacity potentials we use modifications of various alignment procedures implemented in **ToPLign** [11]. **ToPLign** provides procedures to compute several MODES of alignments, global, local, and so called free–shift alignments. The latter do not penalize gaps at the beginning and the end of the resulting alignment and prove to be most useful for threading as the involved sequences often have quite different lengths. Depending on the gap scoring function different ALGORITHMs are used for the optimization: for general gap penalties the ALGORITHM of Needleman and Wunsch [13] requires a quadratic number of memory cells and cubic number of execution steps, for linear gap penalty functions there is a simple algorithm having quadratic time complexity, and for affine gap penalties with costs for opening (gap insertion) and for extending a gap (gap elongation) proposed by Fitch and Smith [7] we use the quadratic time Gotoh–type algorithm [9].
All combinations of MODES and ALGORITHMS mentioned above are implemented in **ToPLign** as slight variations of the following recurrence:

$$\begin{aligned} D_{i,j} &= \mathrm{MAX}(D_{i-1,j-1} + \mathrm{match}(i,j), R_{i,j}, C_{i,j}) \\ R_{i,j} &= \mathrm{Max}_{k<j}(D_{i,j-k} - g_R(i,j,k)) \\ C_{i,j} &= \mathrm{Max}_{k<i}(D_{i-k,j} - g_C(i,j,k)) \end{aligned} \tag{1}$$

This recursion defines the maximal score of the alignments of the i– and j–prefixes of two sequences $\mathcal{R}$ and $\mathcal{C}$. The element $D_{i,j}$ is the maximum of the optimal alignment score of the $(i-1)$ and $(j-1)$ prefixes plus the additional cost match(i,j) for (mis)matching i with j, and the score for the i, $(j-k)$ and $(i-k)$, j prefixes decreased by the cost of a gap of length k in the respective sequences. The cost of a gap of length k in the sequences $\mathcal{R}$ or $\mathcal{C}$ at position (i,j) is denoted $g_R(i,j,k)$ or $g_C(i,j,k)$, respectively.

2.3.2 Path and Confidence Contour Maps

In order to evaluate an alignment, **ToPLign** allows for the computation of a so called *path contour map* P. Such a map is an $n \times m$ matrix labeled with the strings to be aligned and contains at position (i,j) the score of an optimal alignment passing through this particular match.
The path contour accounts at position (i,j) not only for the value of an optimal i and j prefix alignment, but also for an optimal continuation of this path up to the end of both (global), of one of the sequences (free–shift), or as far as

the score stays above zero (local alignment). By definition, all positions on the optimal path carry the same (optimal) alignment score.
In **ToPLign**, extending an idea of Goad&Kanehisa [8] for nucleotid sequences, the path contour matrix is computed from two dynamic programming matrices F ("Forward") and B ("Backward") by applying the dynamic programming process twice – first, for the original strings and second, for the reversed strings.
Intuitively, not reliable alignment positions will be surrounded in a path contour map by high–scoring values. Alternatively, there may be parts of the optimal alignment where any alignment that chooses an alternative route but the optimal one would result in a much smaller score. The latter regions of the optimal alignment tend to be biologically more reliable than the alignment positions mentioned first. Therefore, a reliability or confidence of a specific match on the optimal path, or if we like of any match, can be defined as the score difference of an optimal alignment containing the match and the best alignment not containing this particular match. It is obvious that only matches on optimal alignment paths take positive reliability values.
Again the computation of such a confidence matrix C can be accomplished via the dynamic programming machinery accounting for all paths explicitly avoiding the match (i, j) via standard **ToPLign** procedures.

2.4 Using Contact Capacity Score

2.4.1 Match score

The score defined by the contact capacity potential can easily be figured into the computation of optimal alignments with the dynamic programming recurrence. We modify the term for single matches in the recurrences (1) as follows to be a weighted sum of sequence, local structure preferences and contact capacity potential contributions:

$$\text{match}_{\text{s}}(i, j) = \alpha * s(i, j) + \beta * l(i, j) + \gamma * cc(i, j)$$

where $s(i, j)$ is the sequence score of substituting amino acid j of the structure by the i–th amino acid of the sequence according to Dayhoff [6] type substitution matrix D, i.e. $s(i, j) = D_{ij}$. $l(i, j)$ scores the local preference of the i–th amino acid of the sequence to be in the structural environment class $s(j)$ of structure position j according to the assignment described in section 2.1.1, i.e. $l(i, j) = \text{Pss}(i, s(j))$. $cc(i, j)$ denotes the contact capacity score of mapping i to position j, i.e. the energy assigned to amino acid i to have $nc(j)$ contacts, if $nc(j)$ is the number of actual contacts of the amino acid at structure position j, i.e. $cc(i, j) = \text{Pcc}(i, nc(j))$.

α, β, γ are weighting factors relating the different contributions of the scoring system with respect to each other. For the fold recognition experiments with contact capacity potentials ($\gamma = 1$) reported below, we do not use sequence information at all ($\alpha = 0$) and where secondary structure preference is used its weighting β is 1 and 0 otherwise.
The averaging over a window of length $2w+1$ centered at the match in question is also easily accomplished via:

$$\text{match}(i, j) = \frac{\sum_{k=-w}^{w} \text{match}_s(i+k, j+k)}{2w+1}$$

2.4.2 Gap Penalty

To control gap penalties in conserved structural environments we introduce a parameter σ weighting the contribution of gaps in these regions. To be precise, affine gap costs with gap insertion costs gi and gap elongation costs ge are scored – with $g_R(i, j, k) = g_C(i, j, k) = g(k) = gi + ge * k$ (see equation (1) in section 2.3.1) – as:

$$G_C(i, j, k) = G_R(i, j, k) = \begin{cases} \sigma * \ g(k) & \text{if } s(j) \in \{\mathsf{alpha}, \mathsf{beta}\} \\ \quad\ g(k) & \text{otherwise} \end{cases}$$

For the results presented below we use the following parameter settings: σ is either 1 or 10 depending on whether gap weighting for secondary structures is used or not, the gap insertion parameters used with the different potentials are $gi = 10$ (CCP), 20 (SCCP), 40 (DCCP), 80 (ACCP), respectively, gap elongation ge is set to $gi/10$, and the window size w for averaging match scores is always 3.
These values are tuned such that gaps are reasonably penalized: in almost all cases identity alignments are obtained for native combinations and gaps are introduced in non–native combinations. Until 'optimal' settings are unravelled by parametric analysis the above heuristic reflects the setting of gap penalty values used for sequence alignment with Dayhoff match scores (taking into account the respective potential and window averaging).

3 Results

3.1 Importance of Terms in Potential

Not all the terms in our potential function are equally important for protein fold recognition. One can evaluate the quality of the potential function and the importance of each term with the so called Sippl–test, where sequences are threaded through a set of structures without gaps.

A first analysis compares fold recognition rates of the various forms of our contact capacity potentials. Table 5 shows, that the most detailed contact capacity potential gives the best results in the Sippl–test. The difference, though, between the simple contact capacity (plus secondary structure preference) and the most involved angle dependent potential (ACCP) is minor: The improvement is both for all 167 chains and for 139 chains of > 60 residues less than 3% from 86.8% to 89.2% and 94.2% to 97.1%, respectively.

	SSCCP+SSP	CCCP	DCCP	ACCP
all 167 chains (%correct)	86.8	84.9	83.8	89.2
139 chains > 60 residues (%correct)	94.2	92.8	93.5	97.1

Table 5: Results of the Sippl–test for different kinds of CCP: SSCCP+SSP: Secondary structure dependent CCP plus SS preference; CCCP: Conditional CCP; DCCP: Distance dependent CCP; ACCP: Angle dependent CCP.

Another analysis comparing different types of potentials, – summarized in table 6 – shows that the most important term is the contact capacity potential. This conclusion is consistent with the general concept that hydrophobic forces are the major factor in protein stability. Bryant and Lawrence [3] also noted that in their potential the main contribution to the recognition comes from the hydrophobic term. Russell and Barton [16] computed the energy of common contacts in similar structures and found that the number of stabilizing common contacts in similar structures is almost random, thus neglecting the contribution of the specific pairwise contact potentials. This observation encouraged us to try to use various kinds of contact capacity potentials alone for doing the following fold recognition and alignment experiments.

		SSP+CP+CCP	SSP	CPL	CP	CCP	SSP+CCP	SSP+CP
all 167 chains	Z-score	6.87	2.66	3.10	3.46	5.27	5.55	4.31
	%correct	90.4	24.7	27.1	34.9	76.5	81.9	50.6
139 chains > 60 residues	Z-score	7.30	2.89	2.93	3.42	5.67	6.02	4.43
	%correct	95.7	29.5	27.3	36.0	84.9	91.4	55.4

Table 6: Native Fold Recognition in Sippl–test by different energy functions: SS+CP+CCP: consider all the terms (secondary structure, pairwise contact potentials, and contact capacity potentials); CPL: only long-range pairwise contact potentials; CP: only pairwise contact potentials; CCP: only contact capacity potentials; SS+CCP: without pairwise contact potentials; SS+CP: without contact capacity potentials.

3.2 Fold Recognition Experiments

For the simple shuffle–test (both rotation and random permutation) the native (identity) combination is always recognised with standard deviations of more than 3 (data not shown).

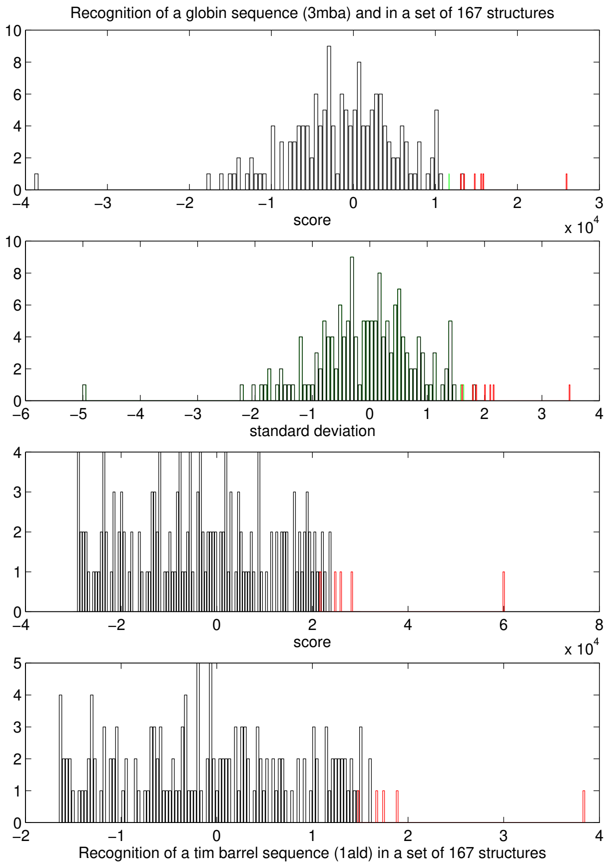

Figure 3: Recognition capability of secondary structure dependent contact capacity potentials (SSCCP)

To perform the most realistic and – from an application point of view – most important threading–test, we wrote a fast program 123D (1–dimensional sequence to 3–D structure), which is now available through the WWW at the URL pages `http://www-lmmb.ncifcrf.gov/~nicka/123D.html` and `http://cartan.gmd.de/ToPLign.html`. In this paper we show the results of two threading experiments: the first with the globin sequence of PDB structure 3mba.pdb, the second with the TIM barrel sequence of 1ald.pdb.
Figure 3 shows some distributions of scores and standard deviations from the average score for 3mba and 1ald for the secondary structure dependent con-

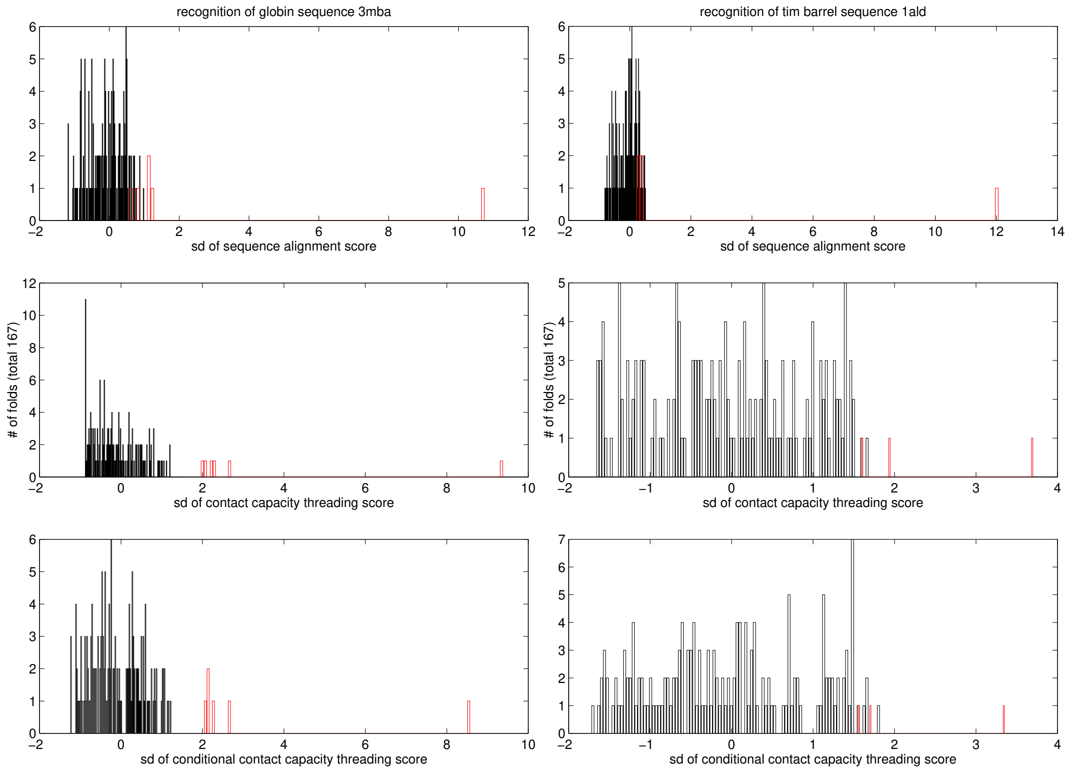

Figure 4: Recognition capability of different potentials: The left column contains the results for 3mba, the right one for 1ald. The upper row shows the distribution of pure sequence alignment scores, which shows a very good discrimation of the native sequence from all the others, i.e. there are no other similar sequences. Some of the next best scores in the distribution do not belong to other globins or TIM barrels and the lowest scoring globin and TIM barrel sequences are at rank 19 and 42, respectively. The middle row contains the distribution of the SS dependent contact capacity scores plus additional sequence score and the lower row the corresponding scores for the conditional contact capacity plus sequence score. Both show that the native combination is not as pronounced as above, but now the related folds, globins or TIM barrels, appear at the top of the list, the globins somewhat separated from the rest of the distribution.

tact capacity potential alone. The native fold is clearly identified and both the other globins (positions 1, 2, 3, 4, 6 and 7, colicin 1colA at position 8) as well as the other TIM barrels (positions 1, 2, 3, 4, and 9) are at the top of the distribution though not separated from the rest. Figure 4 shows the distributions of standard deviations for the same examples for different types of potentials. For these examples the performance of both potentials is comparable. It remains to be shown, whether the more involved conditional potentials can overall improve on the simple CCP's. The most detailed angle dependent potentials (ACCP) show even worse results, which can be explained by the fact that distantly related sequences having the same fold do not preserve such detailed contacts, whereas the contact capacity is much more conserved among similar folds.
We plan to evaluate the performance of the different contact capacity potentials for other recurrent structural motifs and investigate their respective contributions for fold recognition.

3.3 Sequence–Structure Alignments

The use of structural information in the simple and efficiently optimizable form of contact capacity potentials shows significant improvements on alignments of sequences with detectable sequence similarity as well as no significant similarity at all.
Here we discuss two examples of visualizations of optimal and near optimal alignments with path and confidence contour matrices [11] introduced in section 2.3.2. They show that not only the 'correct' alignment can almost perfectly be reproduced but also the number of alternative alignments of similar score is significantly reduced.
Figure 5a shows the path contour matrix for the threading of a globin sequence onto another globin fold using the SS–dependent contact capacity potential without any sequence information. The colour of entries (i, j) in this matrix codes for the score of an optimal threading path passing through this point, i.e. optimal threading alignments containing the match of sequence i onto structural position j. The optimal threading path is shown as white dots. Figure 5b shows the same path contour matrix superposed with the path of the structural alignment, which almost perfectly coincides with the optimal threading alignment shown in black.
Figures 5c and 5d show the reliability contour of the optimal threading, where lighter colours represent higher confidence in the particular match to belong to the optimal alignment. In figure 5d the structural alignment again almost blacks out the optimal path, missing only regions with smaller confidence. The large part in the middle of the matrix with third lowest confidence is actually

a helix, which is displaced one turn in the sequence–structure mapping as compared to the structural superposition. Most probably, in this case, even the structural alignment is wrong.
Figure 6 shows path contour matrices comparing sequence alignment scores with threading scores of the two β–trefoil structures trypsin inhibitor (PDB code 1tie.pdb) and fibroblast growth factor (4fgf.pdb). It can be seen from figures 6a and b that both the local and shift alignments show a quite distorted landscape with the optimal alignments quite far from the structural one. Near the structural alignment path there are no regions which are promising for giving good scores in a sequence alignment.
This situation is greatly improved for the threading scores: The global threading alignment (figure 6c) shows an almost perfect coincidence of the structural with the optimal alignment, however, there is a quite broad area of alignments with almost the same score, i.e. there can be no confidence in this particular optimal alignment even for the threading score. For the shift threading (figure 6d) the optimal regions are much narrower resulting in three different alignment classes with almost the same score reflecting the structural symmetry of the fold. The structural alignment is contained in the quite pronounced second best region, whereas, in this case, it is different from the optimal threading, shown as the region containing the white path.
Figure 7 shows distance maps of these β–trefoil structures (upper left: 4fgf, lower right: 1tie). The corresponding 'aligned distance map', i.e. the rearrangement of rows and columns of the original distance map (top) according to the alignment is shown at the bottom and indicates the quality of the threading alignment.

4 Conclusion

We have derived and tested several modifications of the Contact Capacity Potentials, which reflect the ability of different amino acids to form a certain number of contacts with other residues. The more detailed the potential, the better is the recognition of native folds with Sippl–test, where gaps are disallowed. On the other hand, when gaps are allowed, which is the "real" threading case in modeling protein structures, the detailed potential functions, taking into account either small differences in inter–residue distances, or the distributions of their angular positionings, do not necessarily perform adequately. Investigations of our contact capacity potentials indicates that, compared to pair–wise contact potentials, contact capacity potentials are much more important for the identification of native folds in Sippl–test, which constitutes the standard way for the evaluation of the quality of potential functions. This observation supports developing and applying a fast program for mapping a

1–D sequence to 3–D structure for protein fold recognition, such as the one presented here (123D)[1]. Our program is able to successfully recognize similar structures as demonstrated here in two examples, where the contact capacity score optimized via dynamic programming gives good discrimination in threading with gaps. We have shown that we can improve sequence alignment by using contact capacity potentials almost perfectly reproducing alignments derived from optimal superposition of the associated structures for two examples: two globins with about 20 % sequence similarity and two β–trefoils without significant sequence similarity.

Acknowledgements

We thank Dr. Jacob V. Maizel, for helpful discussions, encouragement and interest. We thank the personnel at the Frederick Cancer Research and Development Center for their assistance. The research of R. Nussinov has been sponsored by the National Cancer Institute (NCI) , DHHS, under Contract No. 1-CO-74102 with SAIC, and in part by by grant No. 91-00219 from the BSF, Israel, and by a grant from the *Israel Science Foundation* administered by the *Israel Academy of Sciences*. The research of R. Zimmer is supported by the German Ministry for Research and Technology (BMBF) under grant number 413-4001-01 IB 301 A/1. Part of this work was done while RZ was visiting scientist at the LMB of the NCI in Frederick. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

[1] A. Bauer and A. Beyer. An improved pair potential to recognize native protein folds. *PROTEINS: Structure, Function and Genetics*, 18:254–261, 1994.

[2] J. U. Bowie, R. Luethy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–170, 1991.

[3] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *PROTEINS: Structure, Function and Genetics*, 16:92–112, 1993.

[1] The program is available via WWW at: `http://www-lmmb.ncifcrf.gov/~nicka/123D.html` and `http://cartan.gmd.de/ToPLign.html`. Experiences and comments are welcome to `nicka@ncifcrf.gov` and `Ralf.Zimmer@gmd.de`

[4] P. Y. Chou and G. D. Fasman. Conformational parameters for amino acids in helical, beta–sheet, and random coil regions calculated from proteins. *Biochemistry*, 13:211–221, 1974.

[5] W. R. David T.Jones and J. M.Thornton. A new approach to protein fold recognition. *Nature*, 358(2 July):86–89, July 1992.

[6] M. O. Dayhoff. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(Supplement 3):345–352, 1978.

[7] W. M. Fitch and T. F. Smith. Optimal sequence alignments. *Proceedings of the National Academy of Sciences USA*, 80:1382–1386, Mar. 1983.

[8] W. B. Goad and M. I. Kanehisa. Pattern recognition in nucleic acid sequences I: A general method for finding local homologies and symmetries. *Nucleic Acid Research*, 10(1):183–194, 1982.

[9] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[10] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3:522–524, 1994.

[11] H. Mevissen, R. Thiele, R. Zimmer, and T. Lengauer. The ToPLign software environment – Toolbox for protein alignment. In *Bioinformatik '94*. Jena, IMB – Institut für molekulare Biotechnologie, 1994.

[12] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structure: Quasi-chemical approximation. *Macromolecules*, 18:534–552, 1985.

[13] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[14] C. Ouzounis, C. Sander, M. Scharf, and R. Schneider. Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from 3d structures. *Journal of Molecular Biology*, 232:805–825, 1993.

[15] O. B. Ptitsyn. How does protein synthesis give rise to the 3d–structure? *FEBS LETTERS*, pages 176–181, 1991.

[16] R. B. Russell and G. Barton. Structural features can be unconserved in proteins with similar folds. *Journal of Molecular Biology*, 244(3):332–350, 1994.

[17] M. Sippl. Calculation of confromational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, 213:859–883, 1990.

[18] R. Thiele, R. Zimmer, and T. Lengauer. Recursive dynamic programming for adaptive sequence and structure alignment. In C. R. et al., editor, *Third International Conference on Intelligent Systems for Molecular Biology*, pages 384–392. AAAI Press, 1995.

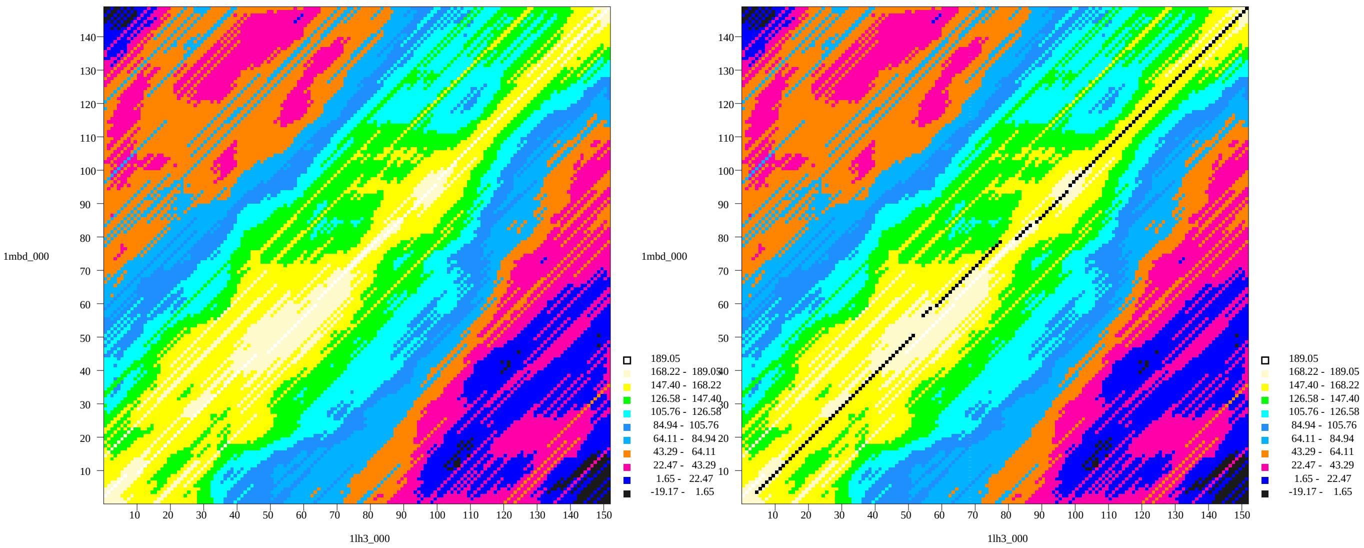

Figure 5a and 5b: Path contour matrix + comparison with structural alignment

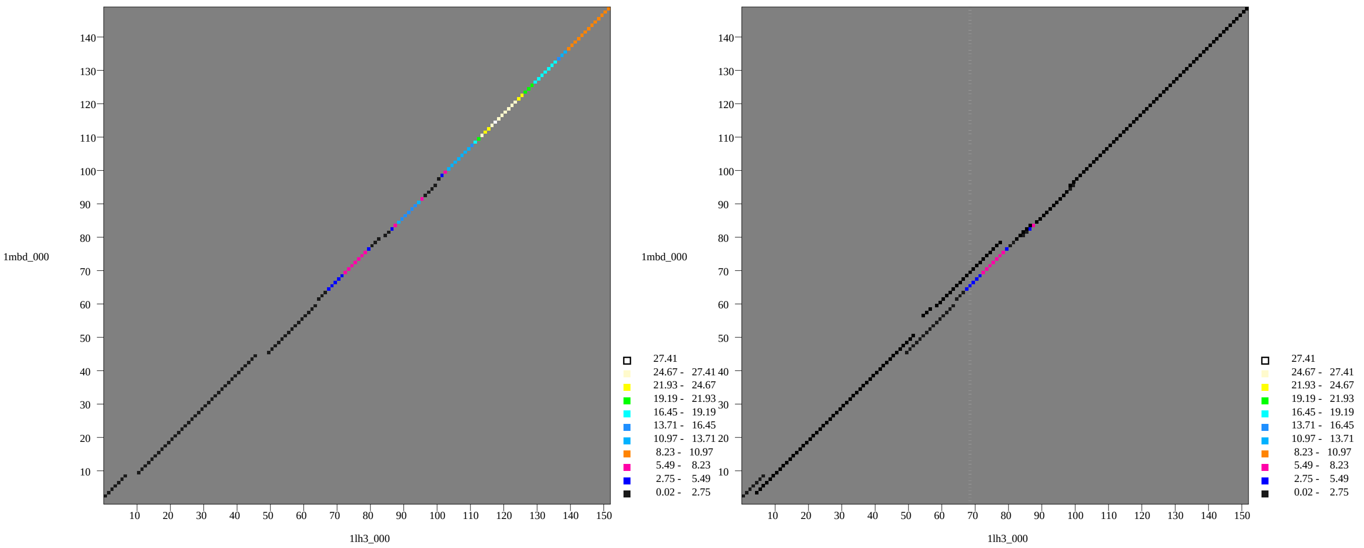

Figure 5c and 5d: Reliability contour matrix + comparison with structural alignment

Figure 5: Contour matrices for the threading of two globins: myoglobin (sperm whale , 1mbd) and leghemoglobin (lupin, 1lh3)

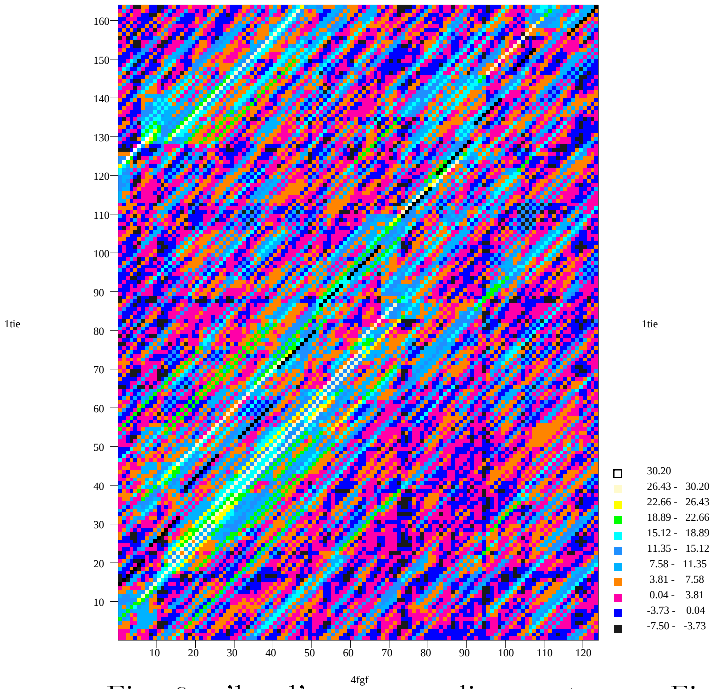

Fig. 6a: 'local' sequence alignment

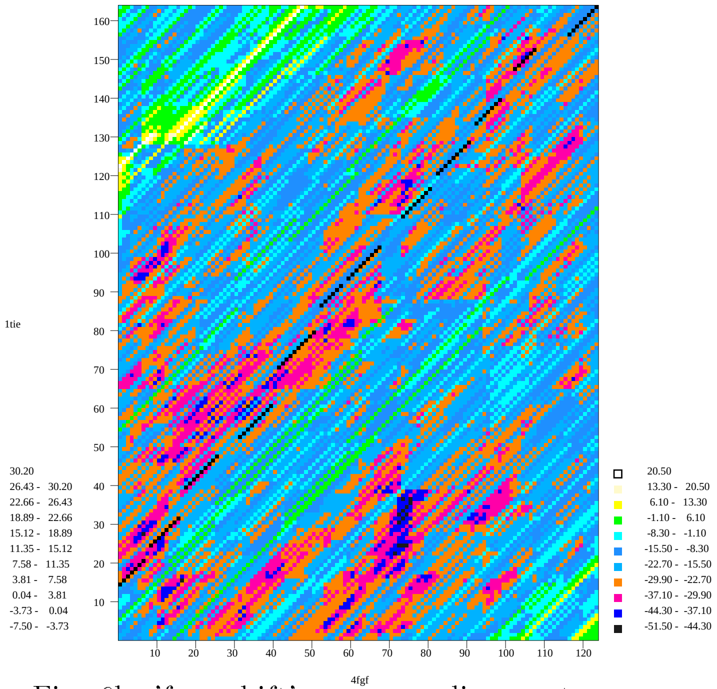

Fig. 6b: 'free–shift' sequence alignment

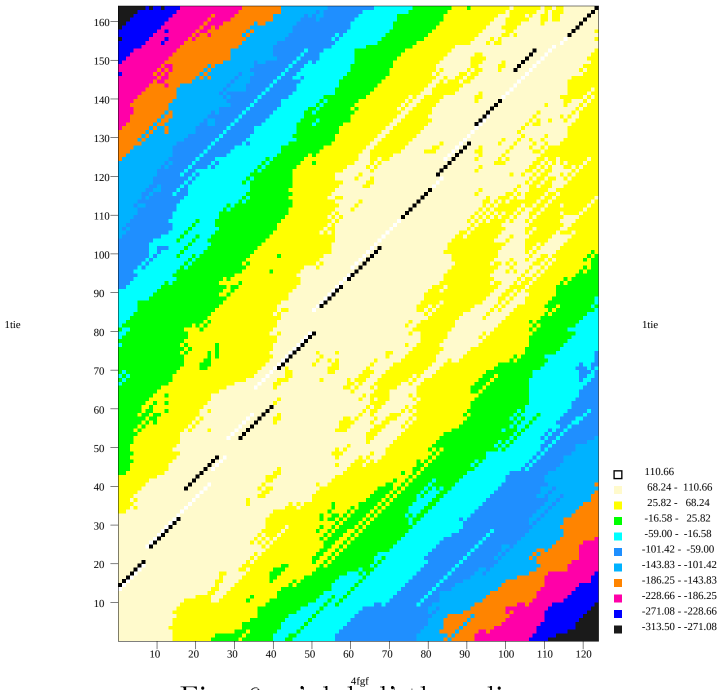

Fig. 6c: 'global' threadings

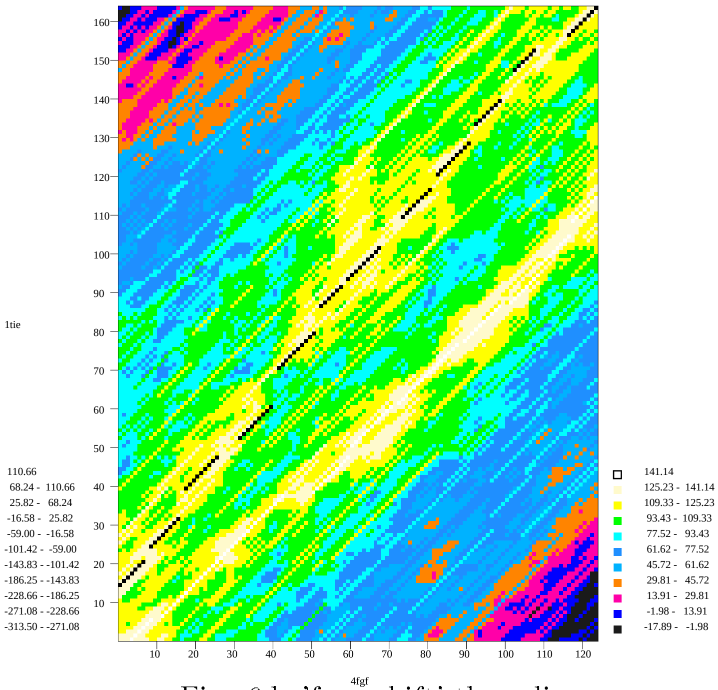

Fig. 6d: 'free–shift' threadings

Figure 6: Path contour matrices with structural alignments for the comparison of sequence alignments with threadings alignments of β–trefoils (trypsin inhibitor (1tie) and fibroblast growth factor (4fgf))

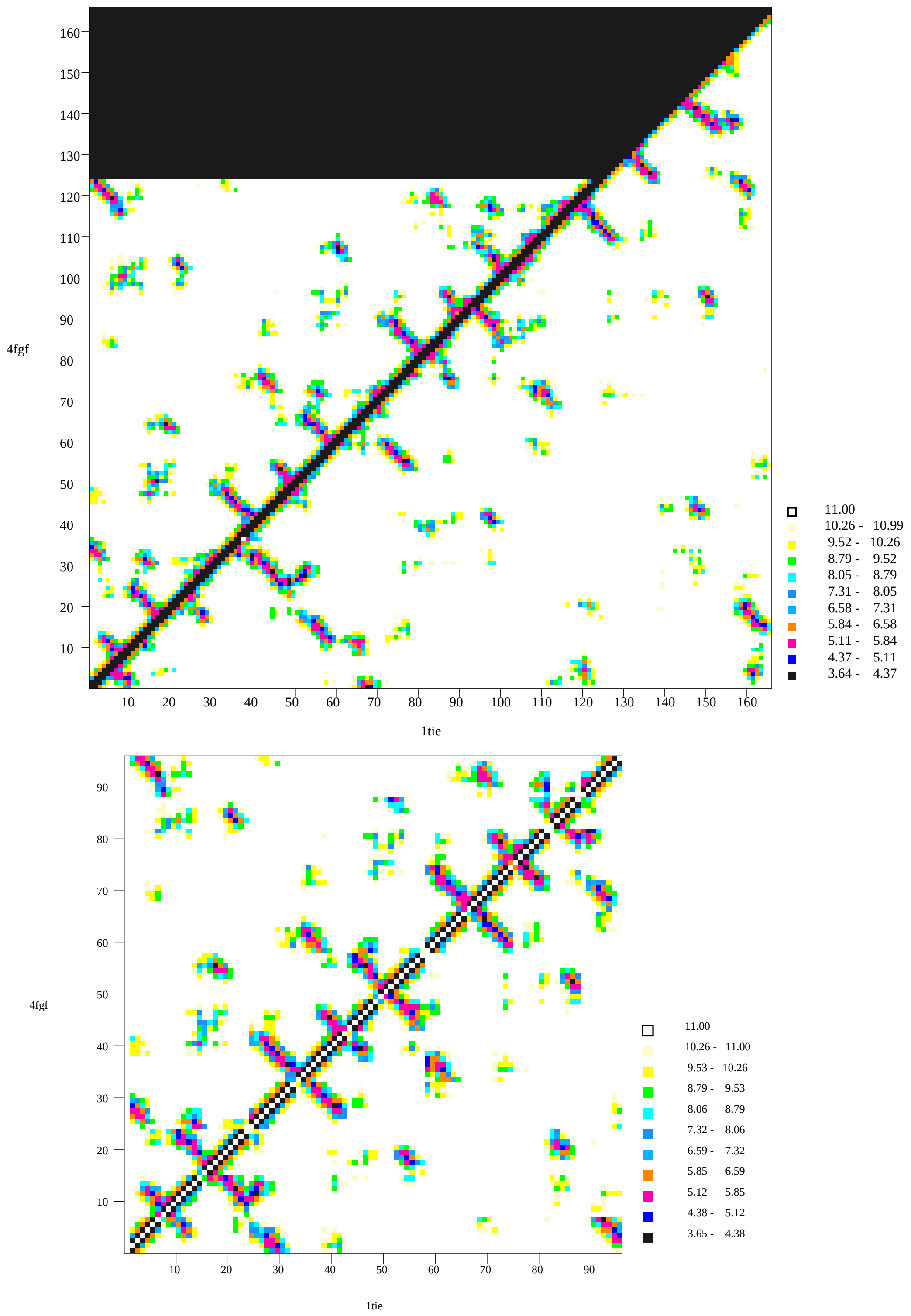

Figure 7: Original (top) and aligned (bottom) distance maps of β–trefoil structures (trypsin inhibitor (1tie) and fibroblast growth factor (4fgf))