Evaluating protein structure descriptors and tuning Gauss integral based descriptors

2005 J. Phys.: Condens. Matter 17 S1523

(http://iopscience.iop.org/0953-8984/17/18/010)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S1523–S1538 doi:10.1088/0953-8984/17/18/010

Evaluating protein structure descriptors and tuning Gauss integral based descriptors

Peter Røgen

Department of Mathematics, Technical University of Denmark, Matematiktorvet, Building 303, DK-2800 Kongens Lyngby, Denmark

E-mail: Peter.Roegen@mat.dtu.dk

Received 1 October 2004, in final form 14 January 2005
Published 22 April 2005
Online at stacks.iop.org/JPhysCM/17/S1523

Abstract

The general development in the natural sciences from relative comparison to absolute description forces the current relative protein structure comparison, based on similarity measures, to be supplemented, or even replaced, by absolute description of each individual protein structure. This paper addresses the question of what should be required from a good set of protein structure descriptors. As an example a Gauss integral based family of protein structure descriptors, that has been shown to successfully classify the geometry of CATH2.4 connected protein domains, is examined. The CATH2.4 domains are here observed to break a symmetry under reversal of the direction of traversal of the protein backbone that general folded tubes possess. It is thus a challenge for any large scale protein or polymer model to explain this broken symmetry.

1. Introduction

The definition and naming of groups is an important part of the natural sciences known as classification. Biology has a long tradition of classification pioneered by Aristotle (384–322 BC) and continued by great naturalists including Carl von Linné and Charles Darwin. In modern times, the huge amount of data in molecular biology has created the need for classification at the molecular level including the classification of proteins [1, 2]. The classification of proteins includes information of various types, the most important being their sequences of amino acids, their three-dimensional native folded structures, and their biological or industrial functions.

In the history of the natural sciences, relative comparison of new samples to the known and best described samples is often replaced by absolute description of all samples. Considering structural classification of proteins, this corresponds to replacing the use of similarity measures by absolute description of each protein structure. Furthermore, to ensure that a descriptor is clear in its objective, available information that is not geometric in nature is ignored, such as sequences of amino acids and the function of proteins when known. To guide the development

of descriptors that can fulfil this task, a list of demands for a good set of descriptors is proposed in section 2. Depending on the focus of interest, the preferred set of descriptors may change. However, when structural classification is the issue, the ability of a set of descriptors to separate folds is central. The only set of protein structure descriptors that has been demonstrated to separate folds consists of 30 descriptors that are based on so-called generalized Gauss integrals [3]. An introduction to these Gauss integrals is given in section 3. In this paper, these Gauss integrals are tuned to meet the proposed descriptor demands as well as possible in section 4. Two of these Gauss integrals were used for protein structure description prior to the work of the author of this paper. These are the writhe used by Levitt [4] and the average crossing number used by Arteca [5]. In comparison, there is much more literature on DNA and these two descriptors [6–10]. In section 5 we discuss which of the descriptors are best for protein structure description. Finally, in sections 6 and 7 it is observed that the set of all CATH2.4 protein domains quantitatively differs itself from general compact folded structures, challenging our understanding of the protein folding process.

2. Demands for a good set of descriptors

Any number that can be associated with a protein structure in a unique way such that it is independent of translation and rotation of the protein structure serves as a descriptor of protein structures. Obvious examples are the volume, the radius of gyration, the moments of inertia around the principal axes, and the accessible surface area that, like other bulk descriptors, describe the protein structure as seen from the outside.

Some descriptors belong naturally to one particular level of the hierarchical division of protein structure into primary, secondary, and tertiary structure. The geometry of the primary structure is given by the length of the protein when imposing a tube picture onto protein structures. At the secondary level, the contents of secondary-structure elements directly and the more geometrical notions of total curvature and total torsion or twist are natural descriptors. Examples of descriptors that measure the tertiary structure are average occurrences of crossing patterns when averaged over all planar projections [11] and a family of generalized Gauss integrals [3, 12].

To propose some demands that a good set of protein structure descriptors, $\{D_1, D_2, \ldots, D_N\}$, should fulfil, denote two protein structures by P_1 and P_2. Each protein structure P defines a feature vector $\{D_1(P), D_2(P), \ldots, D_N(P)\}$ and we equip N-space with the usual Euclidean metric and refer to it as the feature space.

(A) Each descriptor must depend *continuously* on self-avoiding deformations of protein structures.
(B) The set of descriptors must induce a *'nice' pseudo-metric* on the space of protein structures. That is, the usual Euclidean distance in N-space from $(D_1(P_1), \ldots, D_N(P_1))$ to $(D_1(P_2), \ldots, D_N(P_2))$ should be comparable with, e.g., the coordinate root mean square $\mathrm{RMS_c}(P_1, P_2)$ if P_1 and P_2 are highly similar.
(C) The set of descriptors must *separate protein folds*, i.e. '$P_1 \neq P_2 \Rightarrow D_i(P_1) \neq D_i(P_2)$' for some i.
(D) The set of descriptors should be *non-redundant*.
(E) Finally, each descriptor should have a *high signal-to-noise ratio*.

Re point (A). The continuity demand is not fulfilled by the secondary-structure content, by the total torsion or the more general notion of twist [13], and also not by the most probable overcrossing number [14] as argued in [12]. Each of the tertiary-structure descriptors mentioned above depends continuously on self-avoiding deformation of protein structures.

However, some of these tertiary-structure descriptors will make finite discontinuous jumps if a deformation lets the backbone pass through itself. An example is given in section 3.

Re point (B). The number of descriptors, N, is expected to be smaller than the number of degrees of freedom of a protein chain. Many protein configurations thus will have the same feature vector and the metric in feature space gives rise only to a pseudo-metric in the configuration space of protein chains. Point (C) therefore demands that when restricted to natively folded protein structures this pseudo-metric is strong enough to separate folds.

A variety of similarity measures have been applied for protein structure comparison [15–19]. Many similarity measures are based on the coordinate root mean square, $\mathrm{RMS_c}$, or get close to $\mathrm{RMS_c}$ for small deformations of a protein structure. This is the reason that the 'nice' pseudo-metric demand above refers to $\mathrm{RMS_c}$ as the local metric. It is however not beyond doubt that $\mathrm{RMS_c}$ is the best local metric for basing protein structure similarity measures on [20–22] and in general comparison of methods for structural comparison [21–24] is important.

The natural metric to impose on the set of all configurations of a protein is from a geometrical point of view the minimal deformation that can deform the protein from one configuration to another configuration through self-avoiding configurations not counting rotation and translation. Here the size of a deformation is the integral over the entire deformation of the local deformation as measured by $\mathrm{RMS_c}$ or any other local (Riemannian) metric. A way to think of this metric, known as the geodesic distance in the configuration space of the protein, is as the minimal area spanned by the backbone as it undergoes a deformation between two configurations. The area–C^α distance [20] is geometrically similar but does not require deformations to go through self-avoiding configurations.

A deformation that takes the backbone through itself can be made by moving at the most 6–8 carbon alpha atoms 3 Å. The Chebyshev (max norm) used in [20] is for this deformation 3 Å and for a 100-residue protein the $\mathrm{RMS_c}$ is of the order of $\left(8 \times (3\ \text{Å})^2/100\right)^{\frac{1}{2}} \approx 0.85$ Å, significantly below the 'folklore' rule of thumb saying that an $\mathrm{RMS_c}$ below 3 Å implies high structural similarity. In contrast to this, a global self-avoiding deformation distance would be large for this self-intersecting example. Local perturbation metrics such as $\mathrm{RMS_c}$ cannot detect topological changes, but a global deformation metric will probably remain impossible to calculate for many years.

Structural descriptors offer extremely fast all-against-all structural comparison once the descriptor values are calculated for each protein structure. This is since structural comparison is reduced to comparison of points in the feature space, i.e., to comparison of points in Euclidean N-space, where N is the number of descriptors used. In [3] more than 30 000 pairwise comparisons a second were performed using this technique. The metric on feature space gets as close to $\mathrm{RMS_c}$ as possible for self-avoiding deformations of protein structures when a set of structural descriptors is rescaled to meet the 'nice' pseudo-metric demand as well as possible. However, the distance in feature space is free to jump whenever a self-intersecting deformation occurs. Hence, by this rescaling of a set of descriptors, the constructed pseudo-metric reconstructs the global deformation metric as well as possible when given the set of original descriptors and when given the demand of having the usual Euclidean metric in the final feature space.

Re point (C). The demand 'separation of folds' is currently poorly defined as it is not clear how to define protein folds and furthermore it is not clear which cluster methods one should allow in feature space. In a jackknife test 96% of more than 20 000 CATH2.4 domains were correctly identified as belonging to either an existing fold class or to a new fold class [3]. The descriptors used in [3] are based on the number of residues and 29 Gauss integrals. To the author's knowledge, this set of descriptors is the only set that has been demonstrated to separate CATH2.4 domains. The average occurrences of crossing patterns [11] may also separate folds.

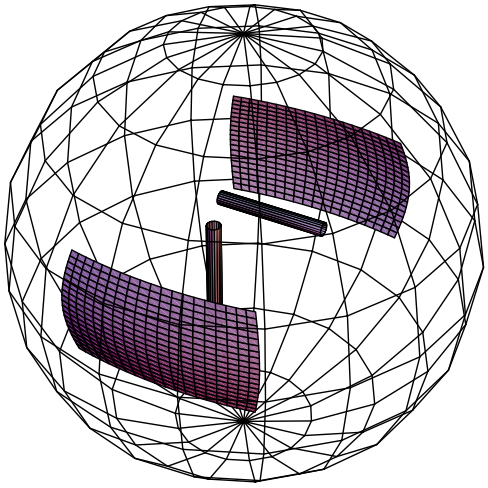

Figure 1. On the unit sphere the filled area corresponds to normals of planes in which the two projected line segments are seen to cross. The two line segments are seen to cross with a probability $|W|$ equal to the filled area divided by the area of the whole sphere when averaged over all directions in space. If the line segment in the front is traversed upward and the rear line segment is traversed from right to left or if both line segments are traversed in the opposite direction a positive crossing is seen and W is positive. Otherwise a negative crossing is seen and W is negative.

(This figure is in colour only in the electronic version)

Re point (D). A descriptor D_N is said to be redundant if there exists a function f fulfilling $f\left(D_1\left(P\right), \ldots, D_{N-1}\left(P\right)\right) \approx D_N\left(P\right)$ for all native protein structures P. In general f could be any type of function, but for a given set of descriptors there are probably certain functions that are natural to consider. The number of residues in a protein structure is assumed to always be known. Hence, the range of any other structural descriptor should be independent of the number of residues.

Re point (E). The idea behind the signal-to-noise ratio of a structural descriptor presented below is to divide the standard deviation of a descriptor on a representative set of protein structures by the average change of the descriptor when small perturbations are applied to the protein structures.

3. Introduction to Gauss integrals

A short introduction to the family of generalized Gauss integrals is included here for the reader's convenience. The natural definition of the writhe for a polygonal space curve μ is

$$I_{(1,2)}\left(\mu\right) = \mathrm{Wr}\left(\mu\right) = \sum_{0<i_1<i_2<N} W(i_1, i_2), \tag{1}$$

where $W(i_1, i_2)$ equals the probability of seeing the i_1th and i_2th line segments cross when averaged over all directions in space times the sign of this crossing, using the usual right-hand rule [12, 25]; see figure 1. An explicit and exact formula for W for given line segments is given in [12, 25]. The writhe has the geometrical interpretation of being the signed average number of crossings seen when averaged over all directions in space. The unsigned average number of crossings seen from all directions is known as the average crossing number and is given by

Table 1. The first column contains the 29 Gauss integrals considered and the pairs mentioned in section 7 are labelled with 'a' to 'f'. The second column contains the correlation coefficient for each Gauss integral and its estimate based on 1053 homology class representatives of CATH2.4. The third column contains the rates at which each Gauss integral and the Gauss integral minus its estimate grow with the number of residues. These growth rates are based on 927 homology class representative domains of CATH2.4 with lengths between 60 and 759 residues. On changing the set of protein domains used, most of these growth rates may change by up to ± 0.1. The fifth column contains the differential geometric signal-to-noise ratio introduced in section 2. The last two columns give the author's choice of descriptors and their relative priorities.

Gauss integral	Correlation coefficient	Growth rate $I/(I-E)$	Signal-to-noise ratio $I/(I-E)$	Choice	Priority
$I_{(1,2)}$	—	0.92	1.91	I	4
$I_{\|1,2\|}$	0.989	1.37/1.00	0.63/0.64	$I-E$	19
$I_{(1,2)(3,4)}$	0.997	1.89/1.01	1.56/1.19	$I-E$	7
$I_{\|1,2\|(3,4)}$ a	0.978	2.26/2.16	1.51/0.79	$I-E$	14
$I_{(1,2)\|3,4\|}$ a	0.989	2.26/2.13	1.46/0.74	$I-E$	16
$I_{\|1,2\|\|3,4\|}$	0.987	2.68/2.13	0.87/1.19	$I-E$	6
$I_{(1,3)(2,4)}$	0.853	0.90/0.88	0.73/0.56	I	17
$I_{\|1,3\|(2,4)}$ b	0.759	1.64/1.68	2.49/1.40	I	3
$I_{(1,3)\|2,4\|}$ b	0.758	1.76/1.69	2.76/1.51	I	2
$I_{\|1,3\|\|2,4\|}$	0.957	2.56/2.17	1.16/1.12	$I-E$	8
$I_{(1,4)(2,3)}$	0.565	1.32/1.48	1.29/1.17	I	5
$I_{\|1,4\|(2,3)}$	0.928	2.43/1.93	1.37/1.03	$I-E$	10
$I_{(1,4)\|2,3\|}$	0.396	1.80/1.99	2.83/1.59	I	1
$I_{\|1,4\|\|2,3\|}$	0.977	2.76/2.36	0.89/1.05	$I-E$	9
$I_{(1,2)(3,4)(5,6)}$	1.000	2.90/1.53	1.59/0.67	$I-E$	18
$I_{(1,2)(3,5)(4,6)}$ c	0.995	1.88/1.96	1.11/0.36	$I-E$	27
$I_{(1,2)(3,6)(4,5)}$ d	0.952	2.65/2.11	1.05/0.84	$I-E$	12
$I_{(1,3)(2,4)(5,6)}$ c	0.995	1.86/2.07	1.07/0.35	$I-E$	29
$I_{(1,3)(2,5)(4,6)}$	0.400	1.69/1.69	0.35/0.33	I	28
$I_{(1,3)(2,6)(4,5)}$ e	0.531	2.09/2.14	0.45/0.42	I	24
$I_{(1,4)(2,3)(5,6)}$ d	0.965	2.81/1.86	1.14/0.88	$I-E$	11
$I_{(1,4)(2,5)(3,6)}$	0.264	1.24/1.19	0.45/0.43	I	23
$I_{(1,4)(2,6)(3,5)}$ f	0.572	1.38/1.23	0.39/0.38	I	26
$I_{(1,5)(2,3)(4,6)}$ e	0.502	2.19/2.09	0.43/0.47	I	25
$I_{(1,5)(2,4)(3,6)}$ f	0.557	2.14/2.10	0.48/0.46	I	22
$I_{(1,5)(2,6)(3,4)}$	0.132	1.80/1.83	0.54/0.53	I	21
$I_{(1,6)(2,3)(4,5)}$	0.907	2.04/2.16	1.11/0.81	$I-E$	13
$I_{(1,6)(2,4)(3,5)}$	0.848	1.19/0.96	0.75/0.54	I	15
$I_{(1,6)(2,5)(3,4)}$	0.466	2.24/2.23	0.60/0.58	I	20

$$I_{|1,2|}(\mu) = \sum_{0<i_1<i_2<N} |W(i_1, i_2)|. \tag{2}$$

The first-order Gauss integral writhe and average crossing number constitute the basic building blocks of a family of generalized Gauss integrals. The following second- and third-order Gauss integrals serve as examples:

$$I_{|1,3|(2,4)}(\mu) = \sum_{0<i_1<i_2<i_3<i_4<N} |W(i_1, i_3)|W(i_2, i_4) \tag{3}$$

$$I_{(1,5)(2,4)(3,6)}(\mu) = \sum_{0<i_1<i_2<i_3<i_4<i_5<i_6<N} W(i_1, i_5)W(i_2, i_4)W(i_3, i_6). \tag{4}$$

Table 1 lists all the Gauss integrals considered in both this paper and in [3]. These integrals are inspired by integral formulae [26, 27] for the Vassiliev knot invariants [28]. The reason for

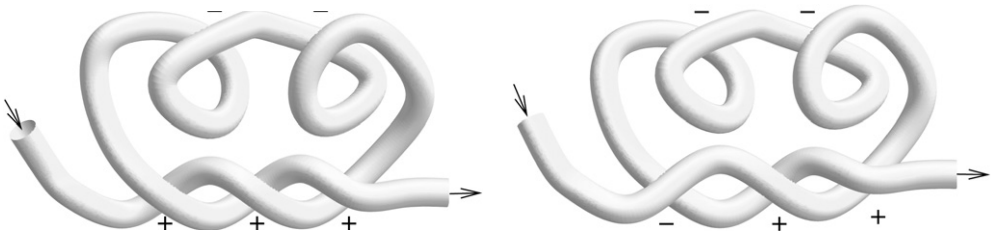

Figure 2. Two tubes with the signs of the crossings seen in this projection. Note that just one crossing is changed from the right-hand side tube to the left-hand side tube.

naming the integrals after Gauss is that the writhe of a smooth space curve, and hereby also the $W(i_1, i_2)$s involved in the above sums, stems from the integral found by Gauss to calculate the number of times that two space curves are linked into each other.

Consider the axis of each tube in figure 2 to be given by a polygonal curve. When the two tubes are squeezed down to almost lie in the plane of this paper, the $W(i, j)$s in the formulae for the Gauss integrals tend either to -1, to 0 (zero), or to +1 as follows. If the line segments i and j are seen to lie apart in the figure, then in the planar limit they will lie apart and in the same plane. The set of directions from which they are seen to cross will diverge to a set of two arcs with measure zero on the unit 2-sphere. That is, $W(i, j)$ tends to zero. However, if the two line segments are seen to cross in the planar projection that they are squeezed into, then in the limit they are seen to cross from all directions on the unit 2-sphere. Hereby, $W(i, j)$ tends to ± 1 depending on the sign of the crossing; see figure 1.

The writhe of the left tube in figure 2 is $I_{(1,2)}(T_{\text{left}}) \approx 3 \times (+1) + 2 \times (-1) = +1$ and that of the right tube is $I_{(1,2)}(T_{\text{right}}) \approx 2 \times (+1) + 3 \times (-1) = -1$. In fact the writhe always jumps in steps of ± 2 when self-intersections occur. In contrast to this, the average crossing number ignores signs of crossings and is ≈ 5 for both tubes. A non-zero term in $I_{(1,2)(3,4)} = \sum_{0<i_1<i_2<i_3<i_4<N} W(i_1, i_2)W(i_3, i_4)$ requires that line segments i_1 and i_2 are seen to cross, that line segments i_3 and i_4 are seen to cross, and that $0 < i_1 < i_2 < i_3 < i_4 < N$. This requirement is equivalent to saying that two crossings are needed and that the line segments corresponding to one of the crossings have to lie downstream of the other crossing. The pair consisting of the two negative crossings, which is the same for the two tubes, is the only pair of crossings fulfilling this demand. Hence, $I_{(1,2)(3,4)}(T_{\text{right}}) \approx I_{(1,2)(3,4)}(T_{\text{left}}) \approx (-1) \times (-1) = +1$. A non-zero term in $I_{(1,4)(2,3)} = \sum_{0<i_1<i_2<i_3<i_4<N} W(i_1, i_2)W(i_3, i_4)$ requires that line segments i_1 and i_4 are seen to cross, that line segments i_2 and i_3 are seen to cross, and that $0 < i_1 < i_2 < i_3 < i_4 < N$. Any pair consisting of one positive and one negative crossing fulfils these requirements on the left tube in figure 2. Thus, $I_{(1,4)(2,3)}(T_{\text{left}}) \approx (+1 + 1 + 1) \times (-1 - 1) = -6$. Similarly $I_{(1,4)(2,3)}(T_{\text{right}}) \approx (-1 + 1 + 1) \times (-1 - 1) = -2$. Note that the discontinuous jump of a generalized Gauss integral on a self-intersecting deformation depends on the entire fold and on the combinatorics of the particular Gauss integral.

4. Tuning Gauss integrals to meet the proposed set of descriptor demands

The number of residues together with the 29 simplest Gauss integrals are shown to separate protein folds in [3]. This corresponds to point (C) in the list of demands in section 2 as previously mentioned. Here, this set of protein structure descriptors is tuned to meet the remaining descriptor demands as well as possible. All these descriptors depend continuously on self-avoiding deformations—see point (A).

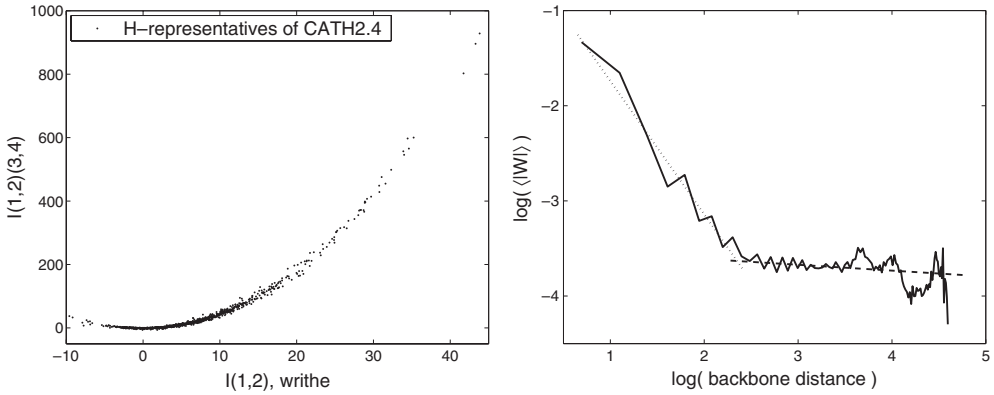

Figure 3. Left: the writhe, $I_{(1,2)}$, versus the Gauss integral $I_{(1,2)(3,4)}$ for a set of 1053 representatives of distinct homology classes of CATH2.4. Right: the logarithm of the average crossing probability, $|W(i_1, i_2)|$, against the logarithm of the backbone distance when averaged over 283 domains with 100 residues. The leftmost point of the graph is averaged over $98 \times 283 = 27\,734$ residue pairs with distance 2 and the rightmost point is averaged over 1×283 residue pairs with distance 99. The floppiness at the tail is due to pure sampling and to the fact that the two ends of a domain are free to be at any relative position in, but usually at the boundary of, the domain. The dotted and the dashed curves indicate almost linear decays for shorter and longer backbone distances.

Re point (D). Some Gauss integrals were found to be correlated in [12]. The clearest correlation is shown on the left-hand side in figure 3. It is not known at this point whether one should just throw away one of these apparently redundant measures or whether the width of this point cloud, shown to the left in figure 3, is a new good structural measure. In order to resolve this, I perform a non-linear Gram–Schmidt orthonormalization-like procedure in which the more complicated Gauss integrals are successively made independent of the less complicated Gauss integrals. The starting point for this procedure is my observation that when averaging over many domains with the same number of residues, the unsigned probability of crossing between two line segments, i.e., $|W(i_1, i_2)|$, of the carbon alpha path decays with the distance along the backbone in a characteristic way. See the right-hand side of figure 3. The average unsigned crossing probability is sampled from more than 24 000 CATH2.4 domains and is approximated by a function of the domain size and of the distance along the backbone. Details can be found in appendix A. This knowledge of the 'average entanglement distribution' along the backbone of protein domains is then used to make estimates of the more involved Gauss integrals when given the values of the simpler Gauss integrals. Details can be found in appendix B. Table 1 shows that 14 of the 29 Gauss integrals may be successfully estimated. At this point it is not known whether each of these 14 Gauss integrals should be discarded or whether subtraction of the estimate, denoted as E, from the value of the Gauss integral, I, gives a good measure, denoted as $I - E$, of the deviation from the average protein structure.

The range of each Gauss integral, or of the Gauss integral minus its estimate, is bounded by the length of the protein as $|I|$ or $|I - E| < C(\#\text{residues})^{\alpha}$, where C and α depend on the Gauss integral only. The αs are estimated as illustrated on the left-hand side of figure 4 and can be found in table 1. From now on, each structural measure is made size independent by division by $(\#\text{residues})^{\alpha}$. The distribution shown in figure 4 on the right is much more independent than the original Gauss integrals, shown in figure 3. However, the level of 'noise' may have been raised considerably by this procedure.

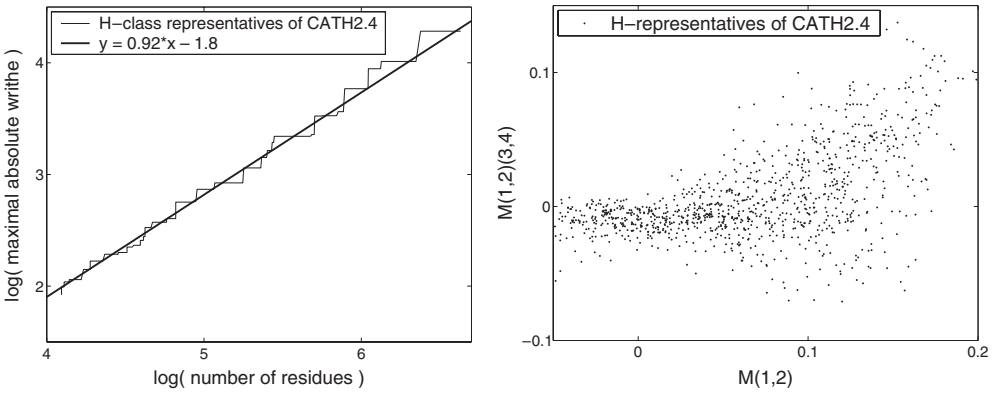

Figure 4. Left: 927 homology class representatives of CATH2.4 are sorted following their number of residues lying between 60 and 759. The logarithm of the maximal absolute value of the writhe of the chains up to any given number of residues is plotted as a function of the logarithm of this number of residues. The straight line is chosen as the line with least mean square distance to the staircase line. Right: the rescaled writhe, $M_{(1,2)} = I_{(1,2)}/(\#\text{residues})^{0.92}$, is plotted together with the measure $M_{(1,2)(3,4)}$ given by $I_{(1,2)(3,4)}$ minus its estimate and divided by $(\#\text{residues})^{1.01}$.

Re point (B). The metric demand is here investigated for each structural measure by itself and not for the entire set of measures gathered in one feature space. This is needed both to define the signal-to-noise ratio below and since it is not known which measures to keep and which to throw away at this step of the analysis. A final descriptor D should fulfil the requirement that $|D(S + \delta S) - D(S)| \approx \mathrm{RMS_c}(S + \delta S, S) \approx |\delta S|$, where S denotes a protein structure and δS is a variation of S that is free of translation and rotation. For each structural measure M a monotonically increasing function ϕ is introduced, such that $D(S) = \phi(M(S))$ fulfils the metric demand, i.e.,

$$\frac{|\phi(M(s)) - \phi(M(S + \delta S))|}{|\delta S|} \approx \frac{\mathrm{d}\phi}{\mathrm{d}M}(M(S)) \left\langle \frac{|\delta M|}{|\delta S|} \right\rangle \approx 1. \tag{5}$$

The average speed at which a measure changes when changing the structure is denoted as $\langle \frac{|\delta M|}{|\delta S|} \rangle$ above and is estimated by performing 100 random variations at the order of 10^{-4} Å for each of 1053 CATH2.4 homology class representatives. Each of these perturbations is made translation and rotation free and preserves neighbouring carbon alpha to carbon alpha distances to first order.

The rescaling function ϕ is chosen as a uniform cubic b-spline, i.e., as a piecewise third-degree polynomial curve. The 50 lowest and the 50 highest values are ignored for each measure to avoid intervals with few data points. The remaining 1053–100 = 953 equations of the form $\frac{\mathrm{d}\phi}{\mathrm{d}M} \langle \frac{|\delta M|}{|\delta S|} \rangle = 1$ are linear in the control points of the spline. The set of control points that gives the minimal root mean square error on this set of linear equations is found. This determines the b-spline ϕ up to an additive constant of integration. The rescaling function is extended linearly outside the interval containing the 953 measure values and finally the additive constant is determined by the claim $\phi(0) = 0$. One of the few rescaling functions that is not approximately linear is shown in figure 5.

Re point (E). When the rescaling functions are applied, all descriptors have the same uniform noise level of 1 (one). Hence, the standard deviation of each descriptor on a representative set of protein structures now defines a signal-to-noise ratio. The set of 1053 CATH2.4 homology class representatives minus the 50 lowest and the 50 highest descriptor values is used as this representative set of protein structures. The signal-to-noise ratios are

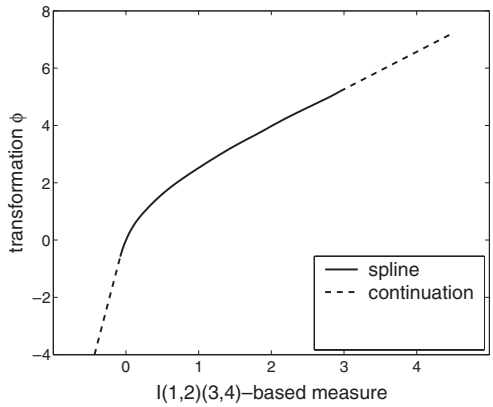

Figure 5. The structural measure $I_{(1,2)(3,4)}/(\#\text{residues})^{1.89}$ is scaled multiplicatively to have standard deviation 1 on a set of 1053 homology class representatives of CATH2.4 and its transformation ϕ as defined in the text.

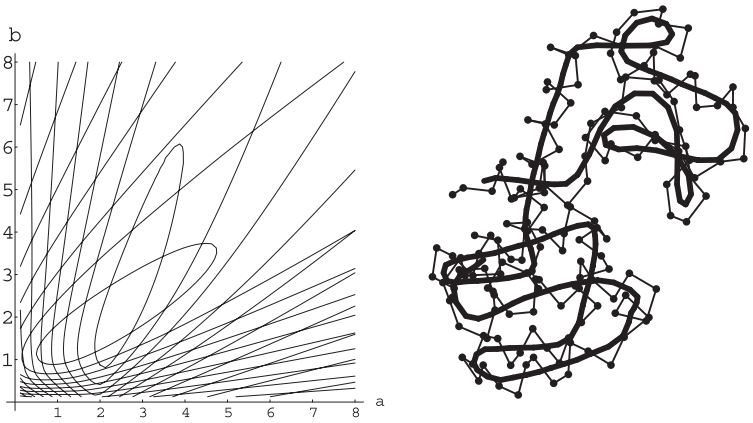

Figure 6. Left: for the mostly alpha protein 7*lzm* and beta protein 2*abx* chain A, the contour lines of the total curvature for varying a and b in the smoothing procedure given in the text are shown. Right: the carbon alpha and the smoothed backbone curve of 7*lsm* for $a = 2.4$ and $b = 2.1$.

reported in table 1 and are of the order of 0.3–3 Å. These ratios are surprisingly low given that the Gauss integrals separate fold classes.

The Gauss integrals depend not only on the curve but also on the curve's tangent image, which here is the curve on the unit 2-sphere connecting the carbon alpha to carbon alpha unit vectors along the backbone. A way to minimize the noise coming from perturbations of the tangents of the curve is to smooth the carbon alpha path locally by exchanging all C_{α_i} with $C_i^{\text{new}} = (C_{\alpha_{i-2}} + a \times C_{\alpha_{i-1}} + b \times C_{\alpha_i} + a \times C_{\alpha_{i+1}} + C_{\alpha_{i+2}})/(2+2 \times a+b)$, where a and b are chosen to minimize the length of the tangent image of the smoothed curve, or in other words to minimize the total curvature of the smoothed curve. The choice $a = 2.4$ and $b = 2.1$ minimizes total curvature for all main types of protein structure, as illustrated in figure 6. The deformation from the carbon alpha curve to the smoothed curve is only performed to the extent that it can be done without self-intersections. This smoothed representation is used in [29]. The above analysis of the Gauss integrals carries over into the case of the smoothed representation and the data are shown in table 2. The two main differences are: (1) the Gauss integrals of the smoothed

Table 2. This table concerns the smoothed backbone. The row |Smooth backbone| concerns the length of the smoothed backbone which is estimated by the number of residues. Otherwise this table is similar to table 1.

Gauss integral	Correlation coefficient	Growth rate $I/(I-E)$	Signal-to-noise ratio $I/(I-E)$	Choice	Priority
|Smooth backbone|	0.974	0.98/0.92	20.6/20.4	$\|\cdot\|-E$	10
$I_{(1,2)}$	—	0.79/—	26.9/—	I	7
$I_{\|1,2\|}$	0.899	1.28/1.01	36.5/35.7	$I-E$	4
$I_{(1,2)(3,4)}$	0.565	1.68/1.56	13.9/16.1	I	15
$I_{\|1,2\|(3,4)}$ a	0.558	2.23/1.88	18.8/21.2	I	12
$I_{(1,2)\|3,4\|}$ a	0.546	2.25/1.92	19.0/21.5	I	11
$I_{\|1,2\|\|3,4\|}$	0.471	2.87/1.50	26.7/44.4	I	8
$I_{(1,3)(2,4)}$	0.466	1.29/1.24	11.0/15.7	I	18
$I_{\|1,3\|(2,4)}$ b	0.222	1.53/1.34	33.3/36.4	I	6
$I_{(1,3)\|2,4\|}$ b	0.161	1.63/1.48	37.8/36.1	I	2
$I_{\|1,3\|\|2,4\|}$	0.222	2.28/0.93	37.5/55.5	I	3
$I_{(1,4)(2,3)}$	0.614	1.36/1.27	16.8/15.9	I	13
$I_{\|1,4\|(2,3)}$	0.612	2.11/1.47	23.1/26.6	I	9
$I_{(1,4)\|2,3\|}$	0.321	1.87/1.76	45.0/38.9	I	1
$I_{\|1,4\|\|2,3\|}$	0.315	2.55/0.94	35.5/57.9	I	5
$I_{(1,2)(3,4)(5,6)}$	0.999	2.74/1.37	9.9/12.5	$I-E$	17
$I_{(1,2)(3,5)(4,6)}$ c	0.978	2.02/1.45	12.0/8.7	$I-E$	27
$I_{(1,2)(3,6)(4,5)}$ d	0.915	1.99/1.51	11.5/10.2	$I-E$	20
$I_{(1,3)(2,4)(5,6)}$ c	0.976	2.24/2.06	11.2/9.1	$I-E$	25
$I_{(1,3)(2,5)(4,6)}$	0.852	1.76/1.57	7.8/9.0	I	30
$I_{(1,3)(2,6)(4,5)}$ e	0.709	1.50/1.44	9.3/9.2	I	24
$I_{(1,4)(2,3)(5,6)}$ d	0.905	2.24/2.01	11.6/10.0	$I-E$	21
$I_{(1,4)(2,5)(3,6)}$	0.762	1.15/1.09	8.7/8.0	I	28
$I_{(1,4)(2,6)(3,5)}$ f	0.441	1.15/0.99	7.8/8.2	I	29
$I_{(1,5)(2,3)(4,6)}$ e	0.777	1.66/1.64	9.9/9.0	I	22
$I_{(1,5)(2,4)(3,6)}$ f	0.227	1.94/1.91	9.3/9.2	I	23
$I_{(1,5)(2,6)(3,4)}$	0.351	1.74/1.76	11.0/10.4	I	19
$I_{(1,6)(2,3)(4,5)}$	0.949	2.42/1.88	12.1/14.2	$I-E$	14
$I_{(1,6)(2,4)(3,5)}$	0.818	1.78/1.75	8.8/8.1	I	26
$I_{(1,6)(2,5)(3,4)}$	0.870	1.80/1.72	13.5/10.2	I	16

backbone are more independent than those of the carbon alpha path and (2) the signal-to-noise ratios have improved and now lie between 7.8 and 45 Å. Hence, the best descriptor, by itself, can be used to detect deformations at the order of 100 Å. The reason for (1) is that helices give very local and numerically very large contributions to all the Gauss integrals. The range of each Gauss integral of the carbon alpha curve is therefore partly given by the amount of helices and partly by the size of the protein. On smoothing the representation, both helices and strands become almost straight line segments and the Gauss integrals can only depend on the topology of proteins, as local geometry is ignored. Interestingly, in [29] it is found that SCOP domains still cluster according to their secondary-structure classification when using the smoothed representation, which in [29] is expected to be caused by the inherent differences in overall packing patterns between protein folds of different types of secondary structure [30].

5. Which measures do we use?

The decision on whether to use a Gauss integral or the Gauss integral minus its estimate is based on the author's studies of the distributions of all the descriptor values and is equivalent

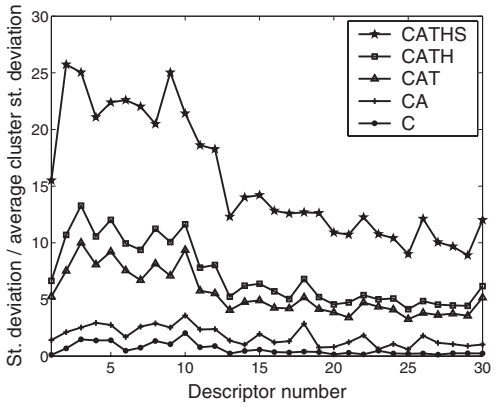

Figure 7. Ordering the descriptors of the smoothed backbone as in table 2, the bio-signal-to-noise ratio defined in the text is shown at the five first levels of the CATH classification system.

to the simple rule: 'If the coefficient of correlation between a Gauss integral and its estimate is greater than 0.9, then the estimate is subtracted'. In tables 1 and 2 the final descriptors are ordered from a differential geometric point of view, namely after decreasing signal-to-noise ratios.

A structural biological rather than a differential geometric test of the descriptive power of the descriptors is based on a data set of 21 026 connected CATH2.4 domains with at most three carbon alphas missing. This data set covers 4, 37, 590, 1054, and 2292 classes at the C, A, T, H, and S classification levels respectively. Figure 7 is based on the smoothed backbone and shows, for each of the five classification levels, the ratio between the standard deviation of the descriptor values and the average standard deviation of the descriptor values within the classes. This bio-signal-to-noise ratio generally decreases with the ordering of the descriptors at the T, H, and S levels. There are even clear linear correlations between this bio-signal-to-noise ratio and the differential geometric signal-to-noise ratio at the sixth and seventh classification levels of CATH that both require very high sequences and structural similarity.

The bio-signal-to-noise ratios of the descriptors of the carbon alpha path are lower than but in the same range as those of the descriptors of the smoothed backbone. On average, the bio-signal-to-noise ratios of the best 12 descriptors have gained 32% at the T level and 38% at the H level through the smoothing procedure, whereas the bio-signal-to-noise ratios of the remaining descriptors are almost unchanged. The calculation times of the Gauss integrals with four and six, respectively, indices depend on the number of residues to the second and third, respectively, powers. With a given application in mind, an obvious idea is to test whether sufficient descriptive power is obtained without using the Gauss integrals with six indices.

6. Are protein domains highly entangled?

A scaling law of the average crossing number, $I_{|1,2|}$, is known for all space curves with a fixed radius. In fact a (closed) space curve of length L and radius R has average crossing number [31]

$$I_{|1,2|} \leqslant \frac{11}{4\pi}\left(\frac{L}{R}\right)^{(4/3)}. \tag{6}$$

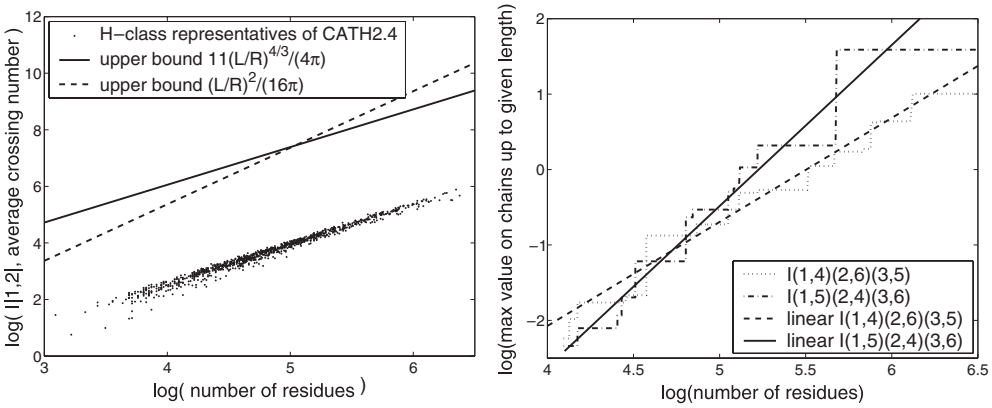

Figure 8. Left: a double-logarithmic plot of the number of residues versus the average crossing number of 1053 homology class representatives of CATH2.4 is shown together with the two upper bounds on the average crossing number given by Buck [31, 32]. The upper bounds are based on setting the distance between neighbouring carbon alpha atoms to 3.8 Å and setting the radius of the backbone to 2 Å. Right: the scaling laws of the Gauss integrals $I_{(1,4)(2,6)(3,5)}$ and $I_{(1,5)(2,4)(3,6)}$ are found, as in figure 4. The two Gauss integrals are found not to obey the same scaling law.

On the CATH2.4 domains this growth rate is found to be 1.37, which is close to the four-thirds power law. However, as illustrated in figure 8 on the right, the constant in the upper bound given by Buck [31] seems high, as also predicted in [32], when applied to protein domains where the constant may be divided by a factor of at least 20. This figure also includes the upper bound $I_{|1,2|} \leqslant \frac{1}{16\pi}\left(\frac{L}{R}\right)^2$ given in [32]. Protein domains are compact and attain thus average crossing numbers following the $\frac{4}{3}$ power law [33]. However, protein domains seem not highly entangled. For example, the high entanglement with distant parts of the curve as seen for highly twisted rubber bands is mostly absent in proteins. In fact the writhe of such a rubber band grows at least with the length of the rubber band and not just with the length to the power of 0.92 (or 0.79 in the case of the smoothed backbone) as found for protein domains. The average crossing number of a closed curve is bounded from below by a so-called crossing number of the knot type, which is the minimal number of crossings seen in any planar projection of any deformation of that particular knot type. Hence, also the absence of 'knots' in most proteins [34], that has the likely explanation of the stickiness of the protein chains [35], points towards a low entanglement of proteins. This raises the question of whether protein domains are maximally entangled, which challenges our understanding of the larger scale mechanisms of the protein folding process.

7. A broken symmetry

Only in the case of the average crossing number is an optimal scaling law known. However, the following symmetry argument shows that at least the Gauss integral $I_{(1,4)(2,6)(3,5)}$ does not grow optimally on protein domains. If the direction of traversal of the carbon alpha path is reversed, then the pairs of Gauss integrals $I_{|1,2|(3,4)}$ and $I_{(1,2)|3,4|}$, $I_{|1,3|(2,4)}$ and $I_{(1,3)|2,4|}$, $I_{(1,2)(3,5)(4,6)}$ and $I_{(1,3)(2,4)(5,6)}$, $I_{(1,2)(3,6)(4,5)}$ and $I_{(1,4)(2,3)(5,6)}$, $I_{(1,3)(2,6)(4,5)}$ and $I_{(1,5)(2,3)(4,6)}$, and, finally, $I_{(1,4)(2,6)(3,5)}$ and $I_{(1,5)(2,4)(3,6)}$ get pairwise interchanged. The remaining Gauss integrals considered here are unchanged under reversal of the direction of traversal. On the set of all self-avoiding polygonal space curves with fixed radius the above pairs of Gauss integrals

share the same scaling law. As illustrated on the right-hand side of figure 8, the scaling laws of the Gauss integrals $I_{(1,4)(2,6)(3,5)}$ and $I_{(1,5)(2,4)(3,6)}$ found on CATH2.4 domains break this symmetry. This symmetry is broken both by the carbon alpha path and by the smoothed backbone. Except for the pair $I_{|1,2|(3,4)}$ and $I_{(1,2)|3,4|}$ the other symmetries appear to be broken, especially by the smoothed backbone. An example of such behaviour is shown by a ball of yarn where the first part of the yarn forms the core and is geometrically different from the last part of the yarn. This suggests that the start and the end of protein chains should not be treated equally when dealing with protein structure prediction and constitutes a benchmark for our understanding of protein folding.

8. Conclusion

With the aim of bringing protein structure comparison from relative comparison of known examples to absolute description of each individual protein structure, I have presented a framework for evaluating protein structure descriptors. The case of a Gauss integral based family of descriptors is then studied in detail. Correlations between half of the original descriptors are identified and removed. Furthermore, the descriptors are made independent of the size of the protein. Through a differential geometric demand I have made the weights of the descriptors at the most indirectly dependent on the set of protein structures used to define them and I have used this to define a signal-to-noise ratio for protein structure description. An algorithm for smoothing the backbone that improves the descriptor signal-to-noise ratios is presented.

The set of all CATH2.4 domains are here found clearly to break a symmetry under reversal of their direction of traversal. It is thus a challenge for any large scale protein or polymer model to explain this broken symmetry.

Acknowledgments

The author would like to thank his colleagues Christian Henriksen and Jens Gravesen for inspiring discussions about parts of this paper.

Appendix A

More than 24000 CATH2.4 domains are used to analyse the average unsigned crossing probability of native folded protein backbones. The average crossing probability $\langle|W|\rangle$ is sampled as a function of the number of residues in the domain $10 < L < 900$ and of the distance along the backbone l counted in residues. Furthermore, for each domain size L and backbone distance l one notes how many times, denoted as $\#eq$, $|W|(L, l)$ appears in the calculation of the average crossing numbers of all these domains. The right-hand side of figure 3 suggests locally approximating $\langle|W|\rangle$ with l^{α}, where α is a function of the size of the domain. This suggests approximating with functions such as

$$e^{S} = e^{(a+b\log(l)+c\log(L)+d\log(l)\log(L))}, \tag{A.1}$$

where S is shorthand for a degree 1 spline in $\log(l)$ and $\log(L)$. The objective is to minimize the total error $\varepsilon\#eq$, where ε is given by

$$\langle|W|\rangle + \varepsilon = e^{S}. \tag{A.2}$$

After division by $\langle|W|\rangle$ and taking the logarithm, we have

$$\log\left(1+\frac{\varepsilon}{\langle|W|\rangle}\right)=S-\log(\langle|W|\rangle). \tag{A.3}$$

On splitting $\log(1+\frac{\varepsilon}{\langle|W|\rangle})$ into its first-order Taylor expansion around $\varepsilon=0$, $\frac{\varepsilon}{\langle|W|\rangle}$, and the remaining higher order term in square brackets, i.e., the identity

$$\log\left(1+\frac{\varepsilon}{\langle|W|\rangle}\right)=\frac{\varepsilon}{\langle|W|\rangle}+\left[\log\left(1+\frac{\varepsilon}{\langle|W|\rangle}\right)-\frac{\varepsilon}{\langle|W|\rangle}\right], \tag{A.4}$$

a linear error estimate is given by

$$\varepsilon=\langle|W|\rangle S-\langle|W|\rangle\log(\langle|W|\rangle)+\left[\varepsilon-\langle|W|\rangle\log\left(1+\frac{\varepsilon}{\langle|W|\rangle}\right)\right], \tag{A.5}$$

where the square bracket, $h=[\varepsilon-\langle|W|\rangle\log(1+\frac{\varepsilon}{\langle|W|\rangle})]$, again holds the higher order terms. Setting $h=h_0=0$ corresponds to replacing $\log(1+\frac{\varepsilon}{\langle|W|\rangle})$ by the first term of its Taylor expansion $\frac{\varepsilon}{\langle|W|\rangle}$ around $\varepsilon=0$. The starting point of the minimization is to find a_1, b_1, c_1, d_1, and ε_1 that minimize

$$\varepsilon_1\#eq=(\langle|W|\rangle S(a_1,b_1,c_1,d_1)-\langle|W|\rangle\log(\langle|W|\rangle)+h_0)\,\#eq, \tag{A.6}$$

for all L and l. Note that this problem is linear in the control points of any spline function, S, used. Now set $\varepsilon_1=\mathrm{e}^{S(a_1,b_1,c_1,d_1)}-\langle|W|\rangle$ and set

$$h_1=\varepsilon_1-\langle|W|\rangle\log\left(1+\frac{\varepsilon_1}{\langle|W|\rangle}\right) \tag{A.7}$$

and repeat the procedure. The general step is thus: find a_n, b_n, c_n, d_n, and ε_n that minimize the root mean square of

$$\varepsilon_n\#eq=(\langle|W|\rangle S(a_n,b_n,c_n,d_n)-\langle|W|\rangle\log(\langle|W|\rangle)+h_{n-1})\,\#eq, \tag{A.8}$$

for all L and l. Set $\varepsilon_n=\mathrm{e}^{S(a_n,b_n,c_n,d_n)}-\langle|W|\rangle$ and set

$$h_n=\varepsilon_n-\langle|W|\rangle\log\left(1+\frac{\varepsilon_n}{\langle|W|\rangle}\right). \tag{A.9}$$

A degree 1 spline with 62 knots (free parameters) in the trapezium given by $12\leqslant L\leqslant 877$ and $1<l<L$ was used in the final approximation. Both the total error and the underlying spline converged under iteration.

Appendix B

Estimates of the Gauss integrals that only involve the absolute values of W can be calculated for each domain length L by use of the average crossing probability distribution along the backbone, $\langle|W|\rangle(L,l)$, established in appendix A. For example,

$$I^{\mathrm{est}_0}_{|1,2||3,4|}(L)=\sum_{0<i_1<i_2<i_3<i_4<L}\langle|W|\rangle(L,i_2-i_1)\langle|W|\rangle(L,i_4-i_3). \tag{B.1}$$

The ratios $R1=\frac{I_{(1,2)}}{I^{\mathrm{est}_0}_{|1,2|}}$ and $R2=\frac{I_{|1,2|}}{I^{\mathrm{est}_0}_{|1,2|}}$ are introduced for obtaining estimates of the Gauss integrals involving the signed W also. Now, e.g., a 'first' estimate of $I_{(1,2)(3,4)}$ is $I^{\mathrm{est}_1}_{(1,2)(3,4)}=I^{\mathrm{est}_0}_{|1,2||3,4|}\times R1\times R1$. These first estimates use the implicit, and wrong, assumption that the signed W is a constant times its absolute value. However, this first estimate is a good estimate up to a multiplicative constant; see figure B.1. From now on, $I^{\mathrm{est}}_{(1,2)(3,4)}$ etc denote $I^{\mathrm{est}_1}_{(1,2)(3,4)}$ times the best possible multiplicative constant. The coefficients of correlation

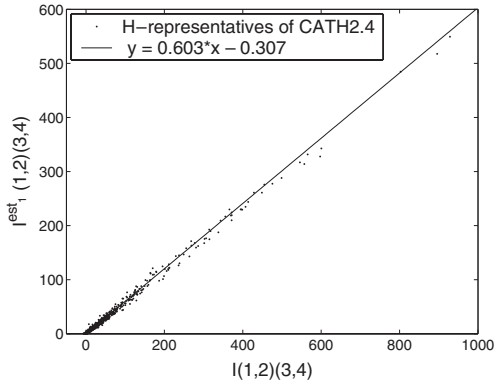

Figure B.1. The Gauss integral $I_{(1,2)(3,4)}$ versus its 'first' estimate $I^{\mathrm{est}_1}_{(1,2)(3,4)}$ defined in the text.

between the Gauss integrals with four indices and their estimates are contained in table 1. From these correlation coefficients and visual inspection, seven out of the twelve estimates are found to be good. Note that the average crossing number, $I_{|1,2|}$, shows only little deviation from its estimate and hence from being a function of the length of the domain.

The estimates of the third-order Gauss integrals are not close except for the case of $I_{(1,2)(3,4)(5,6)}$ and to some extent also for $I_{(1,2)(3,5)(4,6)}$, $I_{(1,2)(3,6)(4,5)}$ and $I_{(1,3)(2,4)(5,6)}$. However, there is more information for building estimates on, namely all the second-order Gauss integrals. Renaming some indices in the product

$$I_{(1,2)}I_{(1,3)(2,4)} = \sum_{0<i_1<i_2<N} W(i_1,i_2) \sum_{0<i_3<i_4<i_5<i_6<N} W(i_3,i_5)W(i_4,i_6), \qquad \text{(B.2)}$$

it is clear that all terms corresponding to $i_2 < i_3$ are also present in

$$I_{(1,2)(3,5)(4,6)} = \sum_{0<i_1<i_2<i_3<i_4<i_5<i_6<N} W(i_1,i_2)W(i_3,i_5)W(i_4,i_6). \qquad \text{(B.3)}$$

Hypothetically, there will be a very close relation between $I_{(1,2)}I_{(1,3)(2,4)}$ and $I_{(1,2)(3,5)(4,6)}$ if the integrals involved are independent of sliding their domains of integration up and down the backbone. The best fit of a linear combination of $I_{(1,2)}I_{(1,3)(2,4)}$, $I_{(1,2)}I_{(1,3)(2,4)}$, $I_{(1,2)}I_{(1,3)(2,4)}$, and the previous estimates over 1053 H-class representatives of CATH2.4 is obtained to improve the estimates of the third-order Gauss integrals. Table 1 and visual inspection reveal that 6 out of the 15 third-order Gauss integrals have good estimates. All in all, 14 of the 29 Gauss integrals are found to be closely given by length of the domains and by lower order Gauss integrals.

References

[1] Orengo C A *et al* 1994 *Structure* **5** 1093
[2] Conte L L *et al* 2000 *Nucleic Acids Res.* **28** 257
[3] Røgen P and Fain B 2003 *Proc. Natl Acad. Sci. USA* **100** 119
[4] Levitt M 1983 *J. Mol. Biol.* **170** 723
[5] Arteca G A 1993 *Biopolymers* **33** 1829
[6] White J H *et al* 1988 *Science* **241** 323
[7] Miller D and Benham C 1996 *J. Knot Theo. Ram.* **6** 859
[8] Podtelezhnikov A 1999 *Proc. Natl Acad. Sci. USA* **96** 12974
[9] Fain B and Rudnick J 1999 *Phys. Rev.* E **60** 7239
[10] Coleman B D and Swigon D 2000 *J. Elast.* **60** 173

[11] Røgen P and Sinclair R 2003 *J. Chem. Inf. Comput. Sci.* **43** 1740
[12] Røgen P and Bohr H 2003 *Math. Biosci.* **182** 167
[13] Røgen P and Bohr H 2002 *MAT-Report 2002-14* Department of Mathematics, Denmark Technical University, Lyngby
[14] Arteca C A and Tapia O 1999 *J. Chem. Inf. Comput. Sci.* **39** 642
[15] Eidhammer I *et al* 2000 *J. Comput. Biol.* **7** 685
[16] Taylor W R *et al* 2001 *Rep. Prog. Phys.* **64** 517
[17] Bostick D and Vaisman I I 2003 *Biochem. Biophys. Res. Commun.* **304** 320
[18] Betancourt M R and Skolnick J 2001 *Biopolymers* **59** 305
[19] Carugo O and Pongor S 2002 *J. Mol. Biol.* **315** 887
[20] Falicov A and Cohen F E 1996 *J. Mol. Biol.* **115** 871
[21] May A C W 1999 *Proteins* **37** 20
[22] Wallin S *et al* 2003 *Proteins* **50** 144
[23] Koehl P 2001 *Curr. Opin. Struct. Biol.* **11** 348
[24] Sierk M L and Pearson W R 2004 *Protein Sci.* **13** 773
[25] Banchoff T 1976 *Indiana Univ. Math. J.* **25** 1171
[26] Bott R and Taubes C 1994 *J. Math. Phys.* **35** 5247
[27] Lin X-S and Wang Z 1996 *J. Differ. Geom.* **44** 74
[28] Bar-Natan D 1995 *Topology* **34** 423
[29] Lindorff-Larsen K *et al* 2005 *Trends Biochem. Sci.* **30** 13
[30] Chothia C *et al* 1977 *Proc. Natl Acad. Sci. USA* **74** 4130
[31] Buck G 1998 *Nature* **392** 238
[32] Buck G and Simon J 1999 *Topol. Appl.* **91** 245
[33] Lua R *et al* 2004 *Polymer* **45** 717
[34] Taylor W R 2000 *Nature* **406** 916
[35] Taylor W R and Lin K 2003 *Nature* **421** 25