Journal of Molecular Graphics and Modelling 32 (2012) 75–81

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

A dynamical approach to contact distance based protein structure determination

Andrew Toon [a], Gareth Williams [b,*]

[a] SIM University, School of Science and Technology, Singapore, Singapore
[b] Bioinformatics, Wolfson CARD, The Wolfson Wing, Hodgkin Building, King's College London, London SE1 1UL, United Kingdom

ARTICLE INFO

Article history:
Accepted 12 October 2011
Available online 26 October 2011

Keywords:
Protein folding
Molecular distance geometry problem
Elastic network model
Go model

ABSTRACT

Protein native structure topology based folding dynamics captures many aspects of protein folding. The fact that folding is driven by a potential derived only from residue pairs in native contact, a sparse distance matrix, lead us to postulate this as a solution method to the molecular distance geometry problem. In the standard Go model non-bonded residues move under the influence of a Lennard–Jones potential and consequently folding is slow. In this study we apply a faster quadratic potential Go model to solving the full-atom distance geometry problem, where distance data is based only on residue atoms within 5 Å in the native structure. We show that the method works well when only atomic contact data is known and when a substantial proportion of this contact data is missing. Also, we show that the method can be applied in conjunction with secondary structure prediction schemes to enhance accuracy in cases of missing contact data.

1. Introduction

The protein folding problem remains one of the great unsolved problems in biology and physics. However, the dynamical behaviour of a protein undergoing folding is largely determined by the final fold topology [1]. This has opened up a fruitful area of research using molecular dynamics (MS) simulations based on Go potentials, where the interaction potential is defined by native residue contacts [2–4]. Go models typically adopt a variety of coarse-grained reductions of the protein structure [5] and define bonded node interactions with a quadratic potential and non-bonded interactions with a species of Lennard–Jones (L–J) potential. Recently, we have developed a damped network model (DNM) of protein folding that essentially models all interactions as quadratic potentials [6]. The DNM is inspired by the success of elastic network models (ENM) in describing the local protein oscillations [7–9], allosteric dynamics [10], allosteric coupling [11] and unfolding behaviour [12]. The extension of the ENM to truly global structure transitions is justified by restricting the interaction range with a simple cut-off based on a realistic influence radius [13]. The DNM based simulations are in good agreement with experimentally observed folding and unfolding dynamics.

One advantage of the DNM over standard Go models is in the speed of folding and the ability to model large proteins. The reason behind the faster folding times of the DNM is that the atomic attraction is strong over the range of influence, typically set to 15 Å, whereas the L–J potential is weak for well separated residues. In particular, the L–J potential tails off as the inverse sixth power of the radius and it is unphysical and computationally unfeasible to set the minima at separations beyond ~7 Å (the average separation of Cα atoms for interacting residues) because of the unavoidability of energy blow ups for close atoms. It can be argued that this is a more physically true picture, but we know that hydropathy drives the initial rapid collapse of the protein to a globular conformation and this is certainly not encoded by inter-residue attractive forces. These considerations aside, the DNM captures many aspects of the folding pathway with physical intermediate structures.

The speed and physical content in DNM lead us to apply DNM to the problem of reconstructing protein structures based on interatomic distance data. Effectively, this is an application of DNM to the molecular distance geometry problem (MDGP) [14,15]. The MDGP is simply the problem of finding atomic coordinates based purely on their mutual separations and is applied to NMR based structure prediction when only residue atom contact information is provided. This concrete application of a minimisation problem has attracted a host of researchers and resulted in many algorithms. The first approach was applied to the case where all distances are given, this was based on single value decomposition (SVD) and consequently at the slower end of the speed scale, $O(N^3)$ for N atoms [16,17]. Since then many algorithms dealing with sparse distance data have been developed, such as the EMBED algorithm [18] based on a combination of missing data generation and SVD, the geometric build up algorithm based on quartet building blocks of structure [19]. Other approaches are described in the recent review [20]. These methodologies are fast and effective though they do not capture the physical nature of the folding process that leads

* Corresponding author. Tel.: +44 2088486806.
E-mail address: gareth.2.williams@kcl.ac.uk (G. Williams).

1093-3263/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2011.10.003

to the target structure. We reasoned that as the DNM captures the physical folding pathway it will generate physical fold solutions to the MDGP and offer an effective way to 'fill in' missing data. We first present the DNM formulation and use it to develop a fast full-distance case solution. Then we present a full-atom approach that is driven purely by contacting atom distance data. We show that the method works well with only atomic contact data and extend the methodology to cases where a substantial proportion of this data is missing. Finally, we show that secondary structure data can be incorporated into the methodology to aid folding when there is limited contact data.

2. Damped network model

The DNM describes a network of nodes moving under a quadratic harmonic oscillator potential given by [6–9]

$$V_{ij} = \frac{1}{2}k_{ij}(d_{ij} - d_{ij}^0)^2, \quad (1)$$

where d_{ij}^0 and d_{ij} are the native and actual Cα separations. The spring constants, k_{ij}, define the coupling strength between atoms and are set to zero for atoms outside the influence radius *Rc*, which physically cannot extend beyond ~15 Å. The protein structure (or a chiral transformation thereof) can be defined based on the minimisation of this potential and with native distances restricted to those corresponding to contacting residues. This is the basis behind NMR distance constraint data based structure determination.

The folding pathway simulation is given a detailed exposition in [6]. In brief, the protein is modeled as a Cα chain and set up with a random conformation, $\vec{r}$, and then follows the equations of motion $m\ddot{\vec{r}} = -(\partial V/\partial \vec{r}) - \mu\dot{\vec{r}}$, where a damping term has been introduced to dissipate the kinetic energy ensuring a convergence to a zero energy conformation. The spring constant is a critical parameter in the folding process as it determines the interaction range and strength. The bonded Cα atoms have a separation of 3.85 Å and a correspondingly large spring constant reflecting the rigidity of the bond. The spring constant vanishes outside the influence radius *Rc*. A standard way of fixing *Rc* is through maximising the correlation between the harmonic fluctuations of the network model and the observed positional uncertainty of the atoms, defined as temperature factors associated with the crystal structure coordinates. This leads to values of *Rc* in the region of 10–15 Å [7–9,21,22].

The DNM approach is relatively fast and can be viewed as a method to derive the structure from the inter-atomic distances. That is a solution to the MDGP. There are many approaches to the MDGP as mentioned above in the introduction. Initial efforts focused on deriving structure based on the full distance matrix. This is an interesting minimisation problem but of limited applicability. Real MDGP applications are when there is only local distance data available, as in the case of NMR distance constraints. The DNM can be applied to both these cases however.

3. Full distance matrix MDGP

When the full distance matrix is known we can develop an extremely fast algorithm for solving the MDGP. To speed up the folding we reasoned that establishing a coarse grained globular configuration would aid the later fine structure derivation. We therefore developed an algorithm based on an initial fast skeleton fold. Here, the protein is represented by small subset of evenly distributed residues, which are folded to a given threshold. The folding state is defined by the value of $D = \sqrt{\sum_{ij}((d_{ij} - d_{ij}^0)^2/N(N-1))}$. The interaction cut-off, *Rc*, is set arbitrarily high as the nodes are in general well separated. Once the skeleton nodes are folded the residues that lie in between the skeleton nodes are introduced to lie evenly distributed along the vectors joining the skeleton nodes. This configuration then serves as the initial conformation for a denser skeleton fold and eventually the full set of Cα atoms is folded to the desired *D*. For a protein of chain length *N* the skeleton nodes *s*(*i*) with spacing *w* are such that $s(1)=1$, $s(i)=(i-1)w$ for $i = 2, \ldots, \lfloor N/w \rfloor + 1$ and $s(L)=N$, where $L = \lfloor N/w \rfloor + 1$ if $N = \lfloor N/w \rfloor \times w$ and $L = \lfloor N/w \rfloor + 2$ otherwise.

Once the skeleton nodes are folded and *D* falls below a given threshold (1.0 Å in our case) the nodes that lie between the skeleton

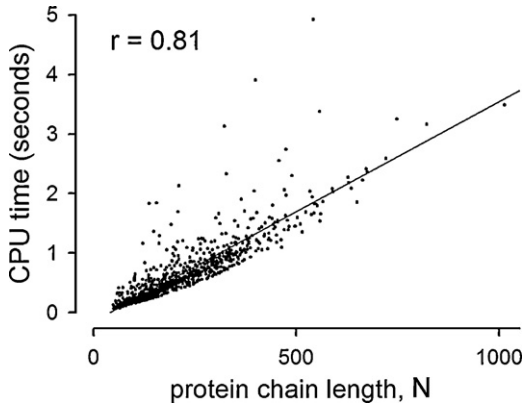

Fig. 1. Folding time scaling of the three stage skeleton neighborhood DNM algorithm. Scatter plot of the average folding CPU time over 5 runs each for 1097 proteins of lengths ranging from 50 to 1000 residues against the protein chain length.

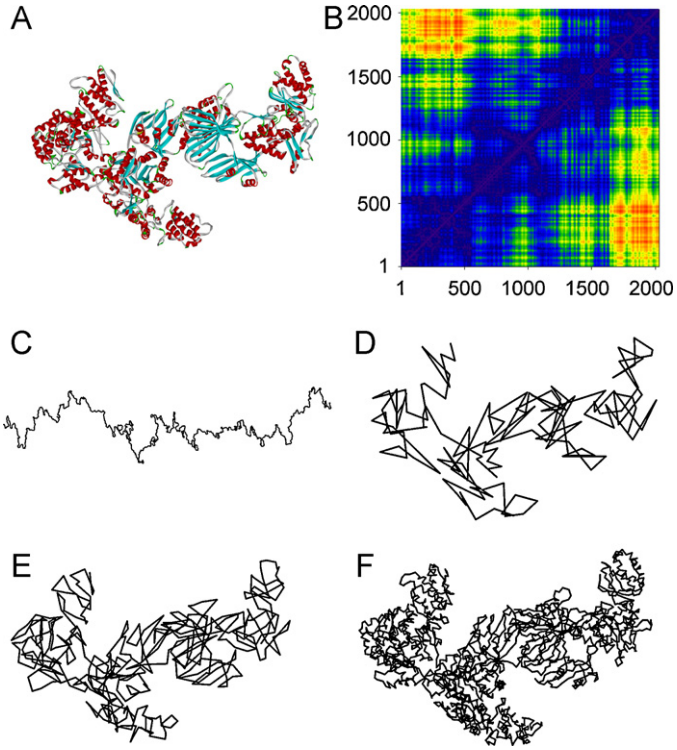

Fig. 2. Folding of fatty acid synthase subunit beta (3HMJ_I) with the three stage skeleton neighborhood DNM algorithm. The crystal structure ribbon representation of the fatty acid synthase subunit beta is shown in A. This is a very large protein consisting of 2033 residues and therefore is a perfect testing ground for the skeleton folding algorithm, which is driven by the full distance matrix, shown as a heat map in B. The initial random non-interacting chain of the Cα nodes is shown in C. The protein is represented by 136 equally spaced skeleton nodes and these are folded with the DNM algorithm resulting in D. The chain is then represented by 407 equally spaced nodes and the resulting fold is shown in E. Finally the full Cα atoms are folded resulting in a structure arbitrarily close to the crystal structure of 3HMJ, F.

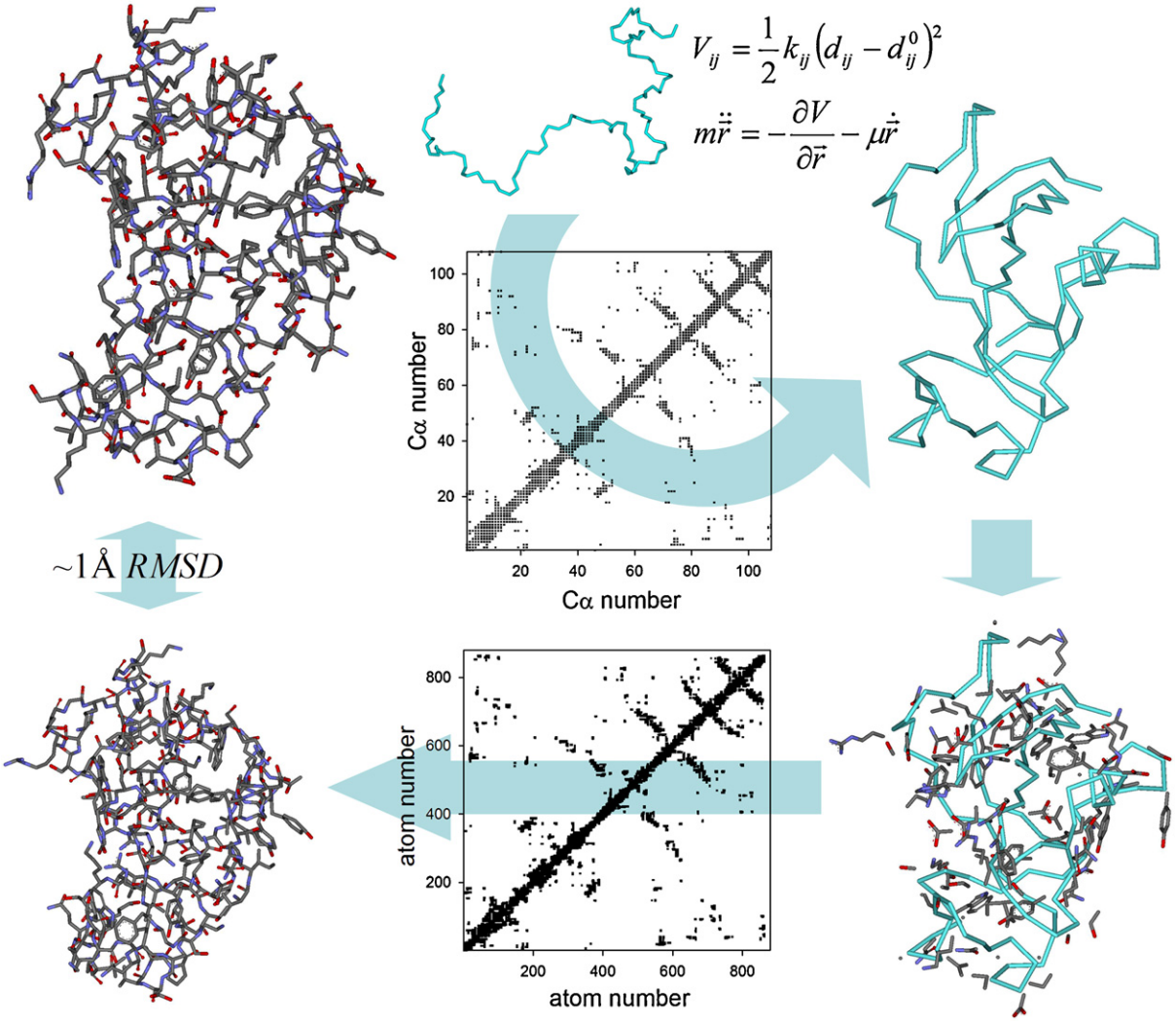

Fig. 3. Schematic of the two stage contact matrix based full atom fold algorithm. The crystal structure (top left) is reduced to a contact matrix for the Cα atoms. This matrix has entries if any atoms from the given residue are closer than 5 Å. A random Cα chain is then folded based on this contact matrix. Once a given tolerance threshold is passed side chain atoms are added to the backbone. The globular full atom conformation is finally folded and scored against the full atom contact matrix. The final structure is in good agreement with the native fold.

nodes are linearly fitted between the skeleton nodes, as follows

$$r_{s(i-1)<s<s(i)} = r_{s(i-1)} + \frac{j - s(i-1)}{s(i) - s(i-1)}\left(r_{s(i)} - r_{s(i-1)}\right)$$
$$j = s(i-1) + 1, \ldots, s(i) - 1 \quad (2)$$

The folded configuration now serves as the starting point for further finer grained skeleton folding.

Within the present context of the DNM we can also restrict the interactions to be between those nodes that are within a given neighborhood in the native conformation. In particular, we order the nodes according to their separation distance and define a set of M nearest neighbors for each node. Folding within the neighborhood we are looking to minimize D for each M nearest neighbors. That is $D(M)$, which is in general smaller than the full D. However, there is a neighborhood size for which small $D(M)$ will imply small full atom D. With $M = 90$ we find that $D(M) < 0.001$ Å rarely corresponds to $D > 1.0$ Å.

To estimate the folding time scaling with protein length we folded a collection of 1097 proteins from the pdbselect25 database. We found the optimum algorithm to consist of the following steps:

1. Establish an initial skeleton fold with $w = 15$ starting from a random non-contacting chain and if the $D < 1.0$ Å threshold is not reached in 5000 folding steps then repeat with another random chain. There was no case for which this process did not terminate in a folded skeleton. This initial step is very fast as the number of nodes is small.
2. Set the skeleton spacing to $w = 15$ and fold to $D < 1.0$ Å.
3. Finally, fold the full set of Cα until the $D < 1.0$ Å threshold is reached or until a given set of iterations is reached. In the last folding stage we set $Rc = 20$ Å.

Each protein was folded five times and the scatter plot of the average run time for each protein against protein length is given in Fig. 1 with a Pearson linear regression coefficient of 0.81. The scatter about the regression line is due to the inherent stochastic nature of the algorithm, where the initial conformation is sampled from an infinite set of random coils. Protein folding is much more efficient with this algorithm and we found that 97% of folding runs resulted in $D < 1.0$ Å. Our folding experiments indicate that a cut-off of 1 Å for D is a valid choice as structures falling below $D = 1$ Å correspond to RMSD coordinate deviation to the native fold of the order of 1 Å $\left(RMS = \sqrt{\sum_i |\vec{r} - \vec{r}_i^0|/N}\right)$. Of course, chirality is invisible to the

Table 1
Folding data for a select set of proteins. Full atom folding runs are presented with exact and fixed atomic distances for contacting atoms. The distance matrix correlation, *D*, is given along with the coordinate RMSD, for the Cα atoms and all atoms, to the native structure. The folding times are given in seconds.

Protein	Chain	Residues	Atoms	Known contact distances				Fixed contact distances			
				D	Time (s)	RMSD(Cα)	RMSD	*D*	Time (s)	RMSD(Cα)	RMSD
1e20	A	185	1457	1.62	62	2.09	2.46	2.15	62	3.18	3.72
1a4c	A	129	977	0.78	19	1.33	1.42	0.87	21	0.95	1.11
1ao3	A	187	1400	0.81	36	0.79	1.01	1.27	96	1.24	1.63
1avg	I	142	1126	0.79	30	0.89	1.03	1.05	32	1.02	1.39
1b0x	A	72	560	2.88	8	5.00	5.03	2.24	6	3.68	4.03
1bft	A	101	804	0.84	26	0.90	1.29	1.45	16	2.36	2.65
1bnl	A	178	1381	1.08	34	1.78	2.07	1.70	34	2.26	2.75
1bou	B	298	2313	1.29	235	1.25	1.59	2.07	101	2.20	2.66
1bvy	F	152	1162	1.27	60	2.05	2.15	1.61	45	2.11	2.60
1bys	A	152	1181	0.80	33	1.10	1.30	1.56	26	1.62	1.93
1c4r	B	178	1363	1.09	31	1.07	1.56	1.25	32	1.66	1.79
1c5f	A	177	1364	0.99	46	1.44	1.72	1.68	33	2.00	2.29
1cbr	A	136	1086	1.09	28	1.41	1.74	1.55	48	1.64	1.94
1col	A	197	1478	1.43	77	1.14	1.90	1.46	28	1.47	1.73
1d2z	A	102	831	1.70	21	2.03	2.56	1.85	36	1.94	2.30
1dyn	A	113	946	0.87	31	0.97	1.44	1.29	37	1.35	1.95
1e2t	A	274	2224	1.27	137	1.32	1.70	2.14	96	3.23	3.35
1ea3	A	157	1209	1.87	34	2.28	2.38	1.41	26	1.36	1.91
1ed1	A	114	893	1.77	43	3.28	3.30	1.35	10	1.93	2.19
1ee6	A	197	1471	0.86	39	1.48	1.66	0.96	33	1.16	1.44
1ej3	A	187	1495	1.07	39	1.23	1.50	2.18	85	2.45	2.65
1exc	A	185	1463	2.25	71	3.37	3.82	1.94	62	2.01	2.48
1eyp	A	212	1610	1.48	78	1.89	2.21	1.82	42	2.07	2.43
1f37	A	109	840	1.40	17	1.92	2.26	1.15	13	1.61	2.09
1fez	A	256	2040	3.71	182	5.17	5.39	4.64	128	7.44	7.99
1fvp	A	231	1873	1.66	132	2.52	2.79	2.52	90	3.38	3.76
1gaq	A	296	2354	2.77	208	4.37	4.43	3.02	133	5.32	5.48
1gd8	A	105	854	1.19	27	2.38	2.67	1.59	17	2.60	2.93
1gnh	A	206	1631	0.89	47	0.94	1.32	1.55	68	1.66	2.03
1hav	A	216	1668	2.20	93	2.93	3.30	2.00	53	2.69	3.11
1hst	A	74	564	1.44	11	2.79	2.79	1.72	7	3.49	3.46
1hyr	A	124	1003	1.16	21	1.38	1.90	1.86	49	2.55	2.99
1i1r	B	167	1363	1.22	67	1.14	1.83	2.12	74	3.19	3.42
1ith	A	141	1062	1.21	44	1.78	1.84	2.01	39	2.51	2.41
1j4x	A	178	1383	1.02	35	1.38	1.54	1.50	24	2.50	2.69
1j90	A	195	1627	1.39	113	1.60	2.20	1.66	91	1.99	2.69
1jd2	N	196	1511	1.24	65	2.23	2.37	2.29	94	2.87	3.55
1jj2	P	95	734	2.72	19	5.69	5.99	2.25	19	4.67	5.41
1jql	B	140	1094	1.44	37	1.77	2.31	1.58	26	1.96	2.47
1jri	E	74	580	1.18	18	1.70	2.05	2.13	6	2.20	3.15
1k0k	A	125	955	0.54	20	0.60	0.86	1.72	24	1.79	2.12
1ktz	B	106	840	0.93	25	0.95	1.58	1.42	13	1.84	2.34
1mhd	A	123	1020	0.96	19	1.32	1.56	1.89	30	2.55	2.97
1nob	A	185	1400	1.66	45	3.38	3.52	1.27	117	2.07	2.27
1p04	A	198	1390	0.92	36	1.56	1.58	1.61	45	2.52	2.61
1pio	A	256	2021	1.56	86	1.58	1.89	1.80	57	2.40	2.74
1qdl	B	195	1548	0.77	65	0.94	1.23	1.41	35	1.07	1.67
1qo3	D	121	988	0.94	21	0.95	1.41	1.24	27	1.30	1.85
1qu0	A	181	1366	0.75	53	1.06	1.22	1.33	25	1.49	1.84
1qu0	A	181	1366	0.83	39	1.04	1.33	1.36	36	1.46	1.71
1tc3	C	51	404	1.79	4	3.27	3.55	1.94	3	2.98	3.33
1tii	D	98	740	1.91	28	2.64	3.13	2.51	15	3.59	4.64
1ukr	A	181	1388	1.05	81	1.14	1.51	1.56	61	1.91	2.24
1xxa	B	72	539	0.84	8	0.56	1.08	1.28	7	1.56	1.97
2pcb	A	294	2370	1.33	137	1.84	2.07	1.96	105	2.65	2.92
3cd2	A	206	1685	1.92	97	2.53	2.79	2.19	80	3.59	3.92
3ygs	P	97	789	1.08	18	1.61	1.46	1.27	9	1.52	1.67

distance matrix and solutions may have to be reflected to obtain native chirality.

To further illustrate the algorithm we present a detailed picture of the folding of fatty acid synthase subunit beta, crystal coordinates deposited under protein data bank accession 3HMJ chain I, comprising 2033 residues, see Fig. 2. Complete folding is achieved in 10 s on a standard personal computer. The protein chain is initiated as a random non-contacting chain Fig. 2A and this is reduced to a set of 136 equally spaced skeleton nodes that are then rapidly folded to $D < 1.0$ Å, Fig. 2B. The skeleton node separation is reduced to 5 residues and the resulting 407 nodes are folded to $D < 1.0$ Å, Fig. 2C. Finally the full Cα atoms are folded to obtain the target structure of 3HMJ.

4. Contact distance matrix based MDGP solutions extended to full atom structure

Practical application of the DNM methodology to MDGP is when only atomic contact data is available. This is the case for NMR structure determination. In simple terms we are given information on the contacts, defined by ∼5 Å proximity, of residue atoms in the

folded structure. The task is then to reconstruct the protein structure. To do this with DNM we adopt a two stage folding strategy. First, we define Cα distances based on the corresponding atomic contact data and fold the backbone accordingly. Then we attach the side chain atoms on to the backbone and initiate a full atom DNM.

Explicitly, we define the full atom coordinates with indices n, m,. . .. and the Cα atoms with indices $i(n), j(n)$,. . .. A pair of residues $i(n)$ and $j(m)$ is deemed to be in contact if $d^0_{nm} < 5$ Å for any constituent atoms n and m. In the Cα folding stage we adopt a potential function $V_{ij} = (1/2)k_{ij}(d_{ij} - \tilde{d}^0_{ij})^2$, where for non-bonded residues $k_{ij} = 1$ and $\tilde{d}^0_{ij} = 7$ Å when $d^0_{i(n)j(m)} < 5$ Å for any n,m and $\tilde{d}^0_{ij} = 9$ Å otherwise. For bonded pairs $k_{ij} = 100$ and $\tilde{d}^0_{ij} = 3.85$ Å . For optimal folding speed the influence radius is set to ∼100 Å for residues in native contact and 9 Å otherwise. The large influence radius models the initial fast collapse of the protein to a globular conformation. The full atom folding stage has an influence radius of 10 Å, but at this stage the influence radius is largely irrelevant due to the proximity of the atoms destined to make contacts.

Once the backbone is folded we populate the side chains based on defined bonded distances and perform a full atom DNM simulation. Independent of conformation, Amino acids of course preserve various distance constraints. These are imposed by the covalent bonding architecture of the protein. Of course, when more is known about the structure than just the contact matrix, for example secondary structure, further distance constraints can be introduced. As above, the spring constant is set to $k_{ij} = 100$ for atom pairs whose separation is constrained by covalent bonding. The potential driving the fold now takes the form $V_{nm} = (1/2)k_{nm}(d_{nm} - \tilde{d}^0_{nm})^2$. Here, $\tilde{d}^0_{nm} = d^0_{nm}$ for $d^0_{nm} < 5$ Å and $\tilde{d}^0_{nm} = 5$ Å otherwise. The spring constant is set to 100 for bonded pairs and unity otherwise. Here, we set $Rc = 10$ Å for atoms in native contact and 5 Å otherwise.

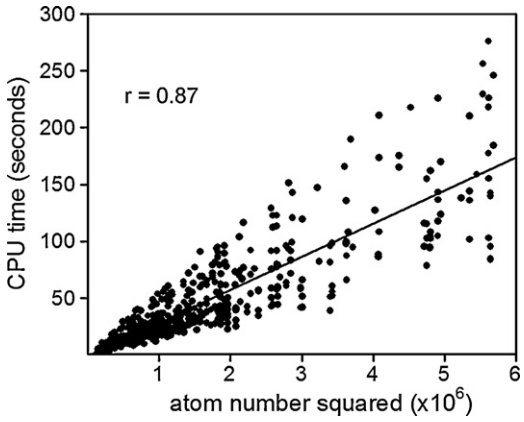

Fig. 4. Folding time scaling of the full atom DNM algorithm. Scatter plot of the folding CPU time for 508 instances for proteins of lengths ranging from 50 to 300 residues against the atom number squared. The linear regression score is 0.87. This is higher than for cubic (0.86) and linear comparison (0.85).

This two stage folding can be thought of as an initial collapse to a globular state followed by refinement. The folding stage is monitored by a Tanimoto score comparison of the native to the DNM folded contacts. Here, $t = N_{ab}/(N_a + N_b - N_{ab})$, where N_a/N_b are the number of native/DNM contact pairs, N_{ab} is the overlap and $0 \leq t \leq 1$. The Cα fold proceeds to $t > 0.6$ and the full atom to $t > 0.80$.

The algorithm is illustrated in Fig. 3 with the example of the 109 residue protein barnase (protein data bank accession 1a2p). In Table 1 we give some example full atom DNM folds. The time complexity of the algorithm is shown in Fig. 4. The time appears to scale with the square of the atom number. In particular, we performed 508 separate full-atom folding runs for proteins of length 50–300

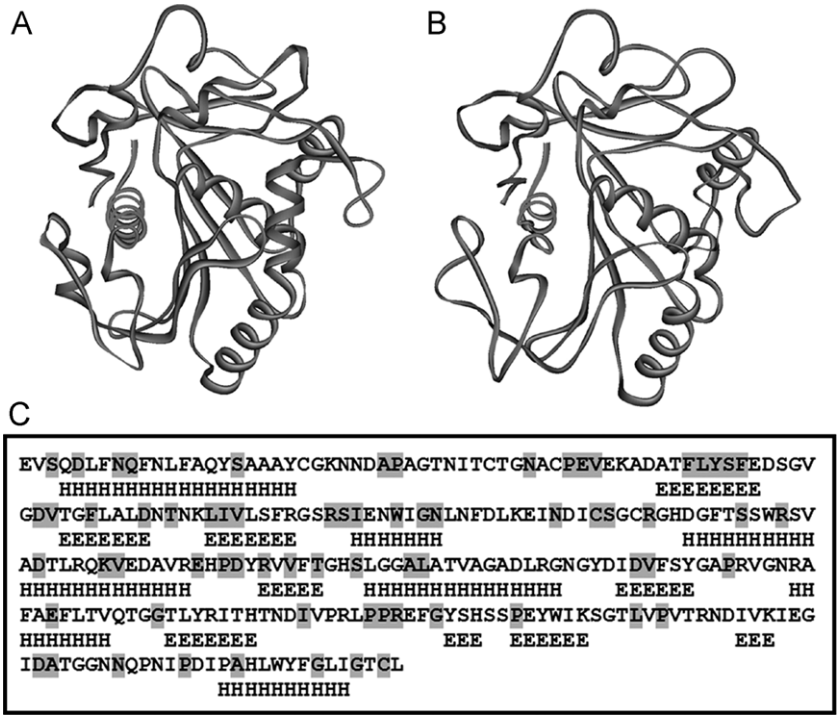

Fig. 5. PSS informed structure determination with missing contact data. An example DNM run for a 269 residue serine lipase with contact data missing for 25% of the sequence. The crystal structure (protein data bank accession 1tib) is shown in A. The predicted structure, shown in B, is very close to the native fold, with Cα and full atom RMSD scores of 2.23 Å and 2.97 Å, respectively. The sequence is shown in C with the secondary structure (H – helix, E – beta sheet) as predicted by the PHDsec prediction module of the PredictProtein web server (www.predictprotein.org) shown under the sequence and the residues for which there is no contact information are highlighted in grey.

residues. The regression score against atom number squared is 0.87 and this is higher than for linear (0.85) or cubic (0.86) scaling.

5. Extension to case of purely binary contact data

In the practical case of NMR structural reconstruction distance data is provided in the form of distance constraints. It is straight forward to recast the DNM scheme to apply to cases where only binary contact data is available. In this case we say that atoms are in contact if separated by 5 Å and give these atoms a predicted separation of 4 Å. Atoms that are not in contact are invisible to each other unless they come within 5 Å when they are repelled. This methodology works remarkably well, converging to structures within a few Angstroms of the native fold. We illustrate with a few examples in Table 1. It is worth noting that the contact matrix Tanimoto similarity score is lower than for the case where distance data is provided, we find 0.6 is an adequate threshold, but the resultant structure is as close to native.

6. The case of missing contact data

The DNM folding pathway captures aspects of the physical folding process. That is, folding intermediates are physically realisable conformations. This observation can be made use of in generating protein conformations where there is missing native contact data. To illustrate this point we set up DNM runs where for a random set of residues we drop all native structure specific contact data, i.e. we run the DNM algorithm with no non-bonded contacts defined for the selected residue set. Remarkably, we find that the fold is recovered in a large number of simulations. In Table 2 we show the folding data for the case of no non-bonded contact information for 25% of the sequence. As the percentage of missing data increases the folding becomes less reliable. However, there are many accurate secondary structure prediction algorithms that can be deployed to aid structure determination. We will see how this can be done explicitly in the next section.

Table 2
Folding data with missing contact data. Folding runs for a set of proteins are given where a random set of residues comprising 25% of the sequence have no non-bonded contact data.

Protein	Chain	Residues	Atoms	$D(C\alpha)$	D	RMSD(Cα)	RMSD
1qu0	A	181	1366	1.66	2.59	1.81	3.05
1ejf	A	110	920	1.55	2.82	1.97	3.82
1ukr	A	181	1388	1.75	2.77	2.08	3.15
1hyr	A	124	1003	1.70	2.32	2.16	3.05
1ejf	A	110	920	1.63	2.59	2.24	3.48
1k0k	A	125	955	2.06	2.82	2.35	3.35
1a4c	A	129	977	1.92	2.80	2.36	3.57
1a4c	A	129	977	2.04	2.74	2.37	3.32
1ith	A	141	1062	2.13	2.55	2.37	2.99
1ejf	A	110	920	1.72	2.73	2.39	3.64
1k0k	A	125	955	1.90	2.61	2.43	3.31
1bnl	A	178	1381	1.98	3.16	2.49	3.61
1tc3	C	51	404	1.64	2.25	2.53	3.47
1qo3	D	121	988	1.92	2.94	2.60	4.15
1c5f	A	177	1364	2.17	3.01	2.63	3.73
1k0k	A	125	955	2.18	3.01	2.64	3.58
1hyr	A	124	1003	1.90	2.51	2.68	3.33
1qdl	B	195	1548	2.40	3.15	2.70	3.77
1nob	A	185	1400	2.26	2.91	2.71	3.58
1a4c	A	129	977	2.16	2.86	2.74	3.84
1ea3	A	157	1209	2.46	3.40	2.75	3.96
1e7k	B	123	954	2.10	2.99	2.75	4.24
1qdl	B	195	1548	2.47	3.22	2.77	3.76
1j4x	A	178	1383	2.34	3.13	2.81	3.88
1col	A	197	1478	2.36	3.17	2.81	3.81
1p04	A	198	1390	2.37	3.13	2.82	3.73
1c5f	A	177	1364	2.37	3.19	2.82	3.96
3ygs	P	97	789	2.35	3.37	2.85	4.30
1ea3	A	157	1209	2.64	3.41	2.88	3.89
1qo3	D	121	988	2.31	3.52	2.90	4.45
1j4x	A	178	1383	2.50	3.32	2.93	4.12
1gnh	A	206	1631	2.23	2.78	2.96	3.61
1qdl	B	195	1548	2.57	3.46	3.00	4.26
1tc3	C	51	404	1.86	2.62	3.00	4.00
1es7	D	86	663	2.19	2.97	3.01	4.28
1jd2	N	196	1511	2.33	3.13	3.04	3.99
1c5f	A	177	1364	2.52	3.43	3.06	4.09
1ej3	A	187	1495	2.59	3.49	3.07	4.16
1qo3	D	121	988	2.08	3.03	3.07	4.30
1ea3	A	157	1209	2.42	2.96	3.08	3.89

7. Secondary structure informed DNM structure prediction

Up to this point we have made do with atomic contact and sequence data. However, there are many knowledge based structure prediction algorithms that can be brought to the aid of the structure solution. Protein secondary structure (PSS) can be predicted based on sequence and local sequence alignment with proteins of known structure. Currently PSS prediction accuracy stands at ~70-80%. It is reasonable therefore to see whether additional distance constraints informed by PSS can be incorporated into the DNM to aid structure prediction specifically in the case where there is missing contact data. Not surprisingly, we found that structure prediction was greatly enhanced when the DNM folding potential incorporated PSS constraints.

There are many PSS prediction servers on line, we chose the PHDsec module of PredictProtein (www.predictprotein.org) [23] and got the secondary structure progression seen in Fig. 5. PSS constraints were built into the folding process by firstly generating initial unfolded conformations with torsion angles restricted to satisfy the PSS types. Folding then proceeds with stringent distance constraints initially on the Cα atoms for the Cα fold and then on the main chain atoms during the full atom fold. Stringency is implemented via a high spring coupling constant, in the same way as for bonded contacts. Of course, more sophisticated distance constraints may be invoked together with torsion angle constraints, but we adopt this economical picture for brevity of illustration. Extensions are straight forward.

To compare the DNM performance with and without PSS constraints in cases of missing contact information we performed 100 folding runs with different randomly selected residues missing non-bonded contact information. The folding runs were performed with and without PSS constraints. We chose the relatively large lipase fold (269 residues, protein data bank accession 1tib) as the testing ground for the methodology. In the case of contact data missing for 25% of the sequence we found sub 2.5 Å Cα RMSD folds in only 8 cases without PSS and 37 cases with PSS. For contact data missing from 40% of the sequence we found one sub 4.5 Å Cα RMSD fold without PSS as against 18 with PSS. It is clear therefore that PSS constraints significantly improve structure prediction. One example run is shown in Fig. 5.

8. Conclusion

Native fold topology based protein dynamics has proved to be a simple and powerful model for studying protein folding and allosteric transitions. In these Go models folding is driven by a potential derived from native atomic contacts defined by atomic proximity. The fact that the fold can be recovered by a dynamical simulation based around native contacts alone lead us to postulate this technique as a solution method to the molecular distance geometry problem. Our formulation of a simple quadratic potential Go model results in faster folding times and therefore is a better candidate for application to the MDGP. In the purely academic

scenario of a full-atom distance matrix our method is comparable to other algorithms in terms of time complexity, scaling linearly with the atom number.

However, in practice the MDGP has to be solved with highly sparse distance information. For example NMR data consists of upper and lower distance bounds on hydrogen atoms within a few ($\sim$5) Angstroms of each other. To this end we have developed a two stage folding approach based on inter-atomic distance data limited to interacting residues. In the first stage a crude backbone fold establishes a globular conformation in rough agreement with the native structure. This scaffold then serves as the starting point for a full-atom fold with side chain atoms attached according to bond length restrictions. We find that solution times scale as the number of atoms squared.

The DNM approach to structure prediction is also shown to work just as well when instead of exact distances being given for contacting residues the only information is whether atoms are in contact. Here, contacting atoms are assigned a predicted separation below the contact threshold. This is more akin to the NMR structure prediction task. Further, the DNM methodology is applied to cases of missing contact information when a substantial proportion of residues have no associated non-bonded contact data. Finally, we show that predicted secondary structure can aid DNM. Here, the initial random conformation is set up with torsion angles satisfying the predicted secondary structure. As the fold progresses stringent distance constraints implemented through large spring constants preserve the secondary structure.

A natural extension of our methodology would be to incorporate knowledge based potentials that are in themselves not of course sufficient to recover the native fold, but can be used in conjunction with sparse contact data.

References

[1] E. Alm, D. Baker, Matching theory and experiment in protein folding, Curr. Opin. Struct. Biol. 9 (1999) 189–196.

[2] H. Taketomi, Y. Ueda, N. Go, Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions, Int. J. Pept. Protein Res. 7 (1975) 445–459.

[3] H. Abe, N. Go, Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins, Biopolymers 20 (1981) 1013–1031.

[4] N. Go, H. Abe, Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation, Biopolymers 20 (1981) 991–1011.

[5] R.D. Hills Jr., L. Lu, G.A. Voth, Multiscale coarse-graining of the protein energy landscape, Comput. Biol. 6 (2010) e1000827.

[6] G. Williams, A.J. Toon, Protein folding pathways and state transitions described by classical equations of motion of an elastic network model, Protein Sci. 19 (2010) 2451–2461.

[7] M.M. Tirion, Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, Phys. Rev. Lett. 77 (1996) 1905–1908.

[8] I. Bahar, A.R. Atilgan, B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, Fold Des. 2 (1997) 173–181.

[9] T. Haliloglu, I. Bahar, B. Erman, Gaussian dynamics of folded proteins, Phys. Rev. Lett. 79 (1997) 3090–3093.

[10] W. Zheng, B.R. Brooks, D. Thirumalai, Allosteric transitions in the chaperonin GroEL are captured by a dominant normal mode that is most robust to sequence variations, Biophys. J. 93 (2007) 2289–2299.

[11] I.A. Balabin, W. Yang, D.N. Beratan, Coarse-grained modeling of allosteric regulation in protein receptors, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 14253–14258.

[12] J.G. Su, C.H. Li, R. Hao, W.Z. Chen, C.X. Wang, Protein unfolding behavior studied by elastic network model, Biophys. J. 94 (2008) 4586–4596.

[13] H. Zhou, Y. Zhou, Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction, Protein Sci. 11 (2002) 2714–2726.

[14] L.M. Blumenthal, Theory and Applications of Distance Geometry, Oxford University Press, London, 1953.

[15] J. Dattorro, Convex Optimization and Euclidean Distance Geometry, Palo Alto, USA, Meboo Publishing, 2005.

[16] G.H. Golub, C.F. van Loan, Matrix Computations, John Hopkins University Press, 1989.

[17] M.J. Sippl, H.A. Scheraga, Solution of the embedding problem and decomposition of symmetric matrices, Proc. Natl. Acad. Sci. U.S.A. 82 (1985) 2197–2201.

[18] G.M. Crippen, T.F. Havel, Global energy minimization by rotational energy embedding, J. Chem. Inf. Comput. Sci. 30 (1990) 222–227.

[19] Q. Dong, Z. Wu, A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data, J. Global Optim. 26 (2003) 321–333.

[20] C. Lavor, L. Liberti, N. Maculan, Encyclopedia of Optimization, 2nd ed., Springer, New York, 2009.

[21] G. Williams, Elastic network model of allosteric regulation in protein kinase PDK1, BMC Struct. Biol. 10 (2010) 11.

[22] I. Bahar, R.L. Jernigan, Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, J. Mol. Biol. 266 (1997) 195–214.

[23] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, Proc. Natl. Acad. Sci. U.S.A. 90 (1993) 7558–7562.