New proteinlike properties of cubic lattice models

Georges Trinquier and Yves-Henri Sanejouand
Laboratoire de Physique Quantique, Centre National de la Recherche Scientifique, UMR No. 5626, IRSAMC, Université Paul-Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France
(Received 13 July 1998)

An analysis of the sequences associated with each of the 60 highly designable structures found in a $3\times3\times3$ cubic lattice model of proteins [H. Li *et al.*, Science **273**, 666 (1996)] reveals that 99% of them belong to a ''neutral island,'' entirely described by a single-mutation walk in sequence space. In each island, five hydrophobic sites are almost perfectly conserved, corresponding to the *unique* five-hydrophobic-residue sequence able to accommodate a three-dimensional hydrophobic core formed by the center of the cube together with the four neighboring centers of face. This happens to reflect the peculiar topologies of the 60 preferred structures. Other properties of the model appear to match specific properties of natural proteins, either structural or evolutionary ones. [S1063-651X(99)02301-6]

PACS number(s): 87.15.By, 82.20.Wt, 36.20.Fz

Proteins are natural heteropolymers exhibiting unusual and highly specific properties. In the past few years, it has been shown that several of these properties can be captured by simple models, such as two- or three-dimensional (3D) lattice models. In these models, the protein is configured as a chain of beads occupying the sites of a lattice in a self-avoiding way and the energy of a sequence folded into a given conformation is obtained as the sum of interactions between unbound first neighbors in the lattice, the sequence itself being specified by the nature of each monomer along the chain.

Focusing on the *compact* arrangements of a chain of 27 monomers located at all sites of a $3\times3\times3$ cubic lattice, there are 103 346 self-avoiding conformations unrelated by symmetry except reverse labeling [1]. This number is relatively small so that the conformation of lowest energy (if any, i.e., if there is no degeneracy) associated with a given sequence can be determined within reasonable computing effort. Starting from an extended arrangement of such a chain along the infinite cubic lattice, it has been shown that for certain sequences the compact conformation of lowest energy can be reached through a Monte Carlo (MC) process within 5×10^7 steps [2,3]. Since there are roughly 10^{16} different possible conformations, this corresponds to the sampling of a small part only of the conformational space of the polymer. Likewise, only a negligible fraction of the 103 346 compact conformations is sampled during the MC course and the choice of the initial arrangement has little consequence, if any, on the outcome of the simulation [2,3]. Such 27-mer sequences therefore appear to behave like natural proteins, which are able to recover their unique native state in seconds, whereas the huge number of their possible conformations prevents any exhaustive sampling (Levinthal paradox [4]). In these simulations, all the proteinlike sequences proved to bear a pronounced energy gap between the lowest and first excited compact states [2].

Recently, a systematic study of all 27-mer sequences was performed within a dichotomic scheme in which monomers are either polar or hydrophobic [5]. For each of the possible 2^{27} sequences, the native compact state was identified, with its associated energy gap. Out of the 103 346 possible conformations, only 120 (0.1%) appear to be highly preferred. Each of these is characterized by a large ''designability,'' as measured by N_s, the number of sequences preferably folding into this arrangement, i.e., having this conformation as their nondegenerate ground state. For these highly designable structures, N_s is found to range from about 1400 to 3800. Furthermore, these preferred conformations are, on average, more stable than the other compact structures since the mean energy gap of their N_s sequences is much larger than those of the remaining conformations [5]. These results strongly suggest that there are only 120 proteinlike structures in the 27-beads chain model, that is, 120 compact conformations in which many sequences are able to fold within a reasonable amount of time. Such properties of the model seem to be shared by natural proteins since a limited number of folds (1000–10 000) is presumed to exist in protein 3D structures [6,7]. Actually, quite different protein sequences may fold into the same native state. As an example, when all known globin sequences are aligned and compared, only one residue out of about 150, a functional histidine, is found to be conserved [8].

In the present work we would like to seek whether there are other properties these 120 preferred structures share with the known folds of natural proteins. The above simple cubic lattice model is kept in its original parametrization, in which the energy $\mathcal{H}$ associated with a given structure is calculated as the simple sum

$$\mathcal{H}=\sum_{i<j} E_{ij}\Delta(r_i-r_j),$$

where $\Delta(r_i-r_j)=1$ if monomers i and j are close neighbors in the lattice (but not in the chain) and $\Delta(r_i-r_j)=0$ otherwise [5]. The increments E_{ij} depend on the polar (P) or hydrophobic (H) nature of the interacting residue monomers: $E_{ij}=E_{HH}=-2.3$ if monomers i and j are both hydrophobic, $E_{ij}=E_{PP}=0$ if they are both polar, and $E_{ij}=E_{HP}=-1.0$ if they are of different nature. Such parameters were derived from an analysis [9] of the Miyazawa-Jernigan matrix of interresidue contact energies between different types of amino acids [10]. For the sake of efficiency, only 60 different proteinlike structures will be considered here, which corresponds to keeping only those preferred con-

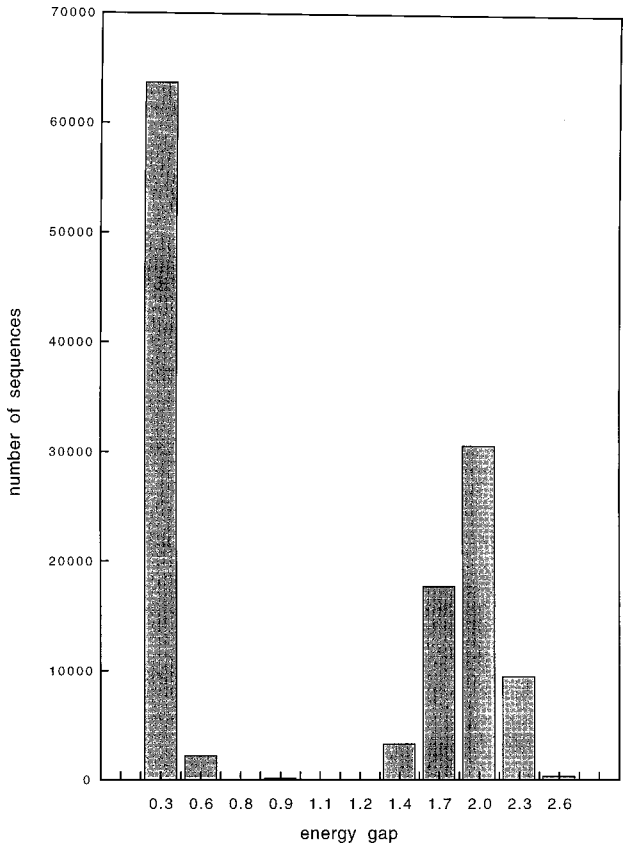

FIG. 1. Histogram of the energy gaps (arbitrary units) of the sequences belonging to the 60 neutral islands.

formations that are unrelated by reverse-labeling symmetry. Although this size reduction is beneficial here and justified, the property that two structures may be identical while having opposite labeling, the first residue in one structure being the last one in the other structure, is certainly not a proteinlike property of the model since a natural protein chain is oriented, due to the asymmetry of the amino acid building block. In fact, the reverse-labeled sequence of a given natural protein could quite well be synthesized. If built from *D*-series amino acids, it should keep the same three-dimensional structure as that of the direct natural sequence. While complete syntheses of all-*D* proteins have been performed [11], to our knowledge no such reverse-labeled proteins have yet been synthesized.

One remarkable property of natural proteins is that when sequences of a given functional protein belonging to two different species are compared, the number of their differences always strongly correlates with the time spent since their common ancestor was living. In the context of the theory of evolution, protein sequences are drifting in a ''neutral island'' of sequence space, as accumulated punctual mutations have no significant effect on the protein function [12]. In other words, starting from a protein sequence in a given species, it must be possible to follow, within the sequence space, a *single-mutation path* ending at the sequence of the same protein in any other species, all proteins along this pathway further keeping the corresponding functionality. In the present model, if a similar property were shared by the N_s sequences of a given proteinlike family or at least by part of them ($N_i < N_s$), then, from any sequence of a given structure, it should be possible to find these N_i neutral-island sequences using a simple step-by-step continuous walk defined as follows. First, all 27 singly mutated sequences are generated. Among them, those who retain the initial structure as their preferred conformation are kept. Next, from each sequence of this subset, 27 singly mutated new sequences are generated and so on until mutation is no longer effective.

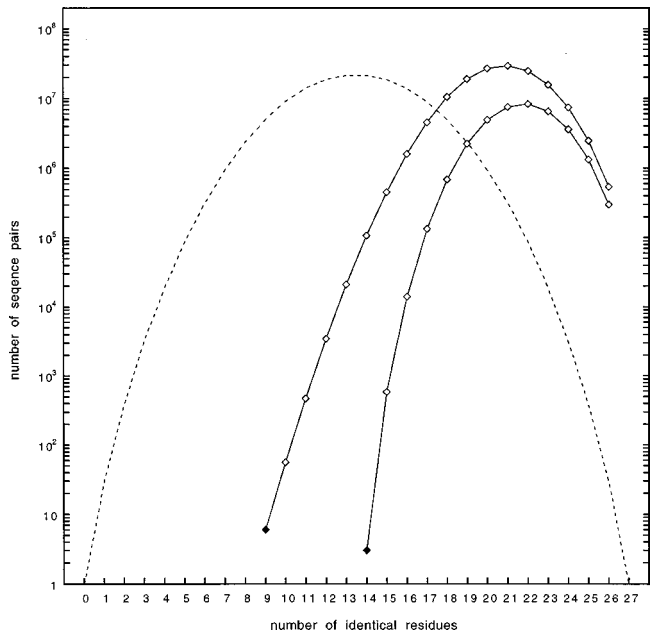

FIG. 2. Sequence identity distributions within all neutral islands (upper curve) and all foldable islands (lower curve). The dashed line refers to the corresponding sets of random sequences.

This protocol was applied to the 60 proteinlike structures of the model. On average, it was found that 99% of the N_s sequences actually belong to the corresponding neutral island. Furthermore, most of the few remaining isolated sequences are only two mutations away from their closer neighbor in the neutral island. However, all sequences belonging to these neutral islands are not expected to behave *a priori* like real protein sequences. Actually, half of them have an energy gap as small as 0.3, as shown in Fig. 1. Outside these 60 islands, the vast majority of sequences are also expected to have small gaps and only one sequence out of 200 000 is found to have a gap larger than 0.6.

According to the above-mentioned criteria [2], small-gap sequences in the neutral islands should not be able to find spontaneously their preferred conformation within a reasonable amount of time. Since, on the other hand, all sequences with a large gap (>1.0) were found to belong to contiguous subsets as defined above, only these sequences should be considered as truly proteinlike. We shall call these subislands the *foldable* islands, as opposed to the *neutral* island. Though the distribution of energy gaps in the neutral islands has a clear bimodal character (Fig. 1), we did not check whether all sequences of the foldable islands actually fold rapidly. This is why both kinds of islands will be considered hereinafter.

In Fig. 2 all pairs of sequences belonging to a same island are compared. These histograms illustrate how some pairs of sequences may have no apparent similarities while being foldable into a same structure (left end of the curves). This

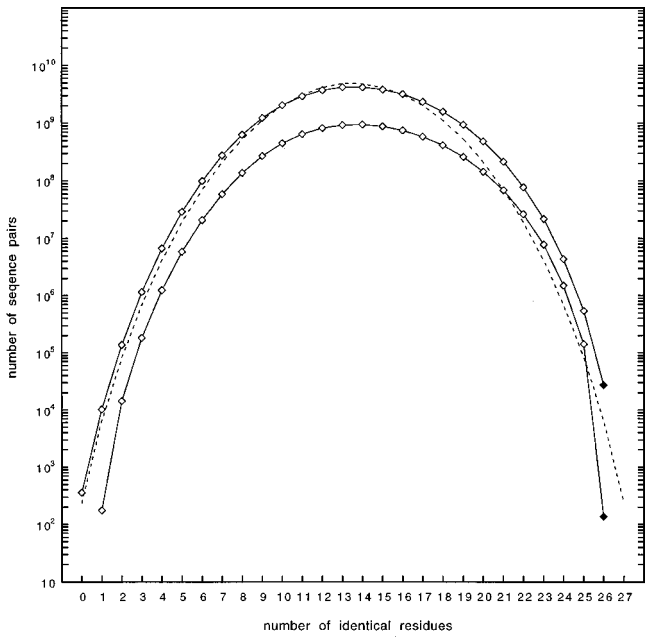

FIG. 3. Sequence identity distributions of all sequence pairs belonging to different islands. Upper curve, neutral islands; lower curve, foldable islands. The dashed curve refers to the corresponding sets of random sequences.

result is in agreement with another well-known property of natural proteins, namely, that two proteins with the same fold may have sequences as different as two sequences picked at random (same fold, weak sequence homology [13]). Sequence comparisons between islands are shown in Fig. 3. As expected, most couples of sequences belonging to different islands (either neutral or foldable) are on average as similar as two random sequences. However, the interesting point here is that there exist numerous cases in which two sequences differ by only one residue while belonging to two different neutral islands or foldable islands (right end of the curves). In the case of natural proteins, this would mean that two sequences of significant similarity may not fold necessarily into the same structure (different fold, strong sequence homology). No such pair of natural protein sequences is presently known, but it may be considered as an open question, as illustrated by the recently met [14] ''Paracelsus challenge'' proposed by Rose and Creamer, whose purpose was to design proteins of different folds with more than 50% sequence identity [15]. On the other hand, the conversion between two foldable islands via a single mutation is not without recalling the postulated conformational switch of prion proteins [16]. These points will be developed elsewhere in more detail.

For sequences of both kinds of island, the probabilities of finding a hydrophobic residue at each type of lattice site are listed in Table I. If monomers located on the edges (edge centers) are most variable, those located in the bulk (cube center) are always hydrophobic. More strikingly, in the foldable islands, the four monomers located on the faces (face centers) that are not linked to cube center [17] are also perfectly conserved as hydrophobic. Together, these five conserved hydrophobic residues form the hydrophobic nucleus of the 60 structures, quite analogous to the conserved hydrophobic cores of proteins [20] or to the critical core of Lau and Dill's model [21].

TABLE I. Probability (%) of finding a hydrophobic residue at a particular site of the cubic lattice for the 127 399 sequences of the neutral islands and 61 371 sequences of the foldable islands. The face-center sites are analyzed separately, depending on whether or not they are linked to the cube center.

Site type	Neutral island	Foldable island
Cube center	100	100
Face center (nonbonded)	98.5	100
Face center (bonded)	89.8	91.2
Edge center	41.5	38.8
Corner	6.6	0.2

Plotting the average energy gap of neutral-island sequences as a function of their total number of hydrophobic residues (Fig. 4) reveals that foldable sequences are more likely to have small counts of hydrophobic monomers. In each foldable island, there is only one five-hydrophobic-residue ($5H$) sequence of large gap (2.3) [22], which suggests that the reason why only 60 proteinlike structures were previously found should originate in simple topological arguments. Indeed, the 60 large-gap $5H$ sequences are the only ones with *five* hydrophobic residues lined up in such a way that a *unique* compact conformation among the 103 346 possible ones can accommodate the corresponding hydrophobic core. Similar conclusions were reached from a recent analysis of the neighborhood vectors of the preferred structures (each of these 60 vectors, the ith component of which is the number of neighbors of the ith monomer, was found to be unique [23]) in line with the statement of Tang and Wingreen that the preferred structures are ''atypical'' ones [24]. One consequence of this state of things is that most proteinlike properties are expected to be only slightly sensitive to changes in the details of the model. Indeed, when different energy increments are used (e.g., $E_{HH}=-3.0$, $E_{PP}=0$, and $E_{HP}=-1.0$), again the same 60 sequences are found to be

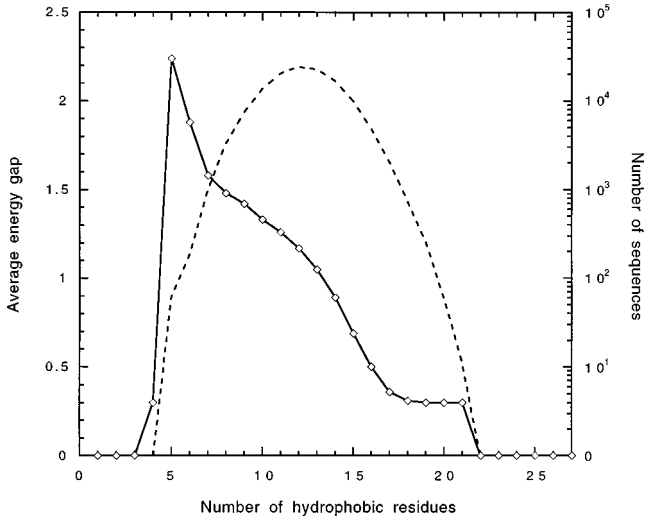

FIG. 4. Average energy gap (arbitrary units) against number of hydrophobic residues for the sequences of the 60 neutral islands. The dashed line corresponds to the histogram of the number of hydrophobic residues for all sequences belonging to these islands (right-hand vertical axis).

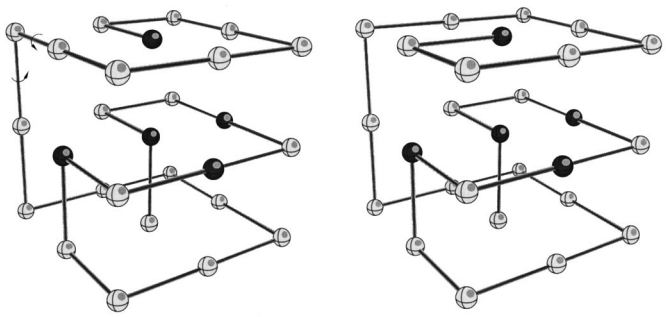

FIG. 5. Example of a bistable structure, analogous to a protein with two functional conformational states.

the only ones with five hydrophobic residues and a large gap. Not surprisingly, however, the sizes of neutral and foldable islands are in this case modified. As an example, for the ''top'' structure of largest N_s, N_i is only 3575 with the above parameters, instead of 3790 with the standard ones [25] or 2306 in the case of the additive potential ($E_{HH}=-2.0$, $E_{PP}=0$, and $E_{HP}=-1.0$).

To fulfill their biological function, many proteins switch from one conformation to another via large-amplitude conformational motions induced by small changes in their environment. In this context, we further considered model sequences exhibiting a degenerate ground state. An analysis of all $5H$ sequences apt to form hydrophobic cores in any compact form indicated that 105 of them select two degenerate preferred conformations. As exemplified for the sequence P**H**PP**H**P**H**P**H**PPPPPPPPPPPPPPPPPP**H**, the two preferred conformations are often similar, differing only in the relative orientation of two domains, here of unequal size (Fig. 5). Such a motion is quite analogous to hinge motions or shear motions well documented in real proteins [26]. As many as 2956 sequences are found in the neutral island around this sequence. Defining here the energy gap as the difference between the lower of the two structures preferred by the $5H$ sequence and the lowest state among all other ones, one finds an average energy gap of 0.79, which is as large as that found in several of the 60 proteinlike families. With an energy gap larger than 1.0, 26% of the sequences of this island are expected to fold rapidly. In Fig. 6, the neutral islands corresponding to the 105 bistable structures are compared to the neutral islands corresponding to the 60 single-ground-state structures discussed above. The two families are of similar sizes, with similar average energy gaps, and one can expect that these 105 bistable structures also exhibit many proteinlike characters.

The heteropolymer considered in the present study does not look like a protein in many respects: It has only 27 residues (instead of 50–5000 in a typical protein) of only two

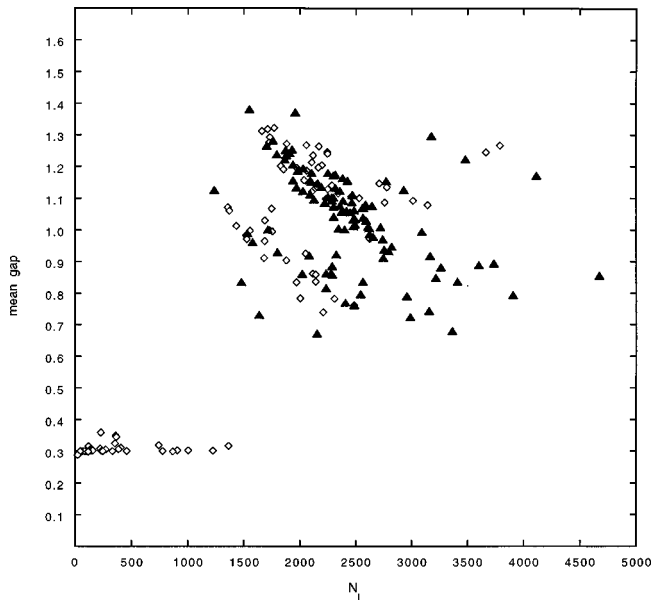

FIG. 6. Average energy gap (arbitrary units) as a function of N_i, the number of sequences in neutral islands. Filled triangles, singly degenerate ground-state islands; empty diamonds, single-ground-state islands, together with a few structures of the low average gap, shown for the sake of comparison.

kinds (while there are 20 different amino acids), with a single bulk one (while nearly half of them are buried in a typical protein), etc. In spite of this, the variety of properties shared by this simple model and by natural proteins suggests that it may carry the essence of their specificities, which is quite encouraging for subsequent modeling.

This $3\times3\times3$ lattice model not only provides structural features such as the limited number of folds, the crucial role of the hydrophobic core, or the existence of two-conformation families, but it seems to reproduce evolutionary properties as well. If large neutral islands of sequences can be explored through single mutations without changing foldability into a given structure, evolutionary shifts into other stable structures can also take place at low mutation expense. In a way, such a behavior is consistent with major concepts underlying macroevolution, as it is presently understood, such as neutralism [12] and punctualism [27]. The following challenges will consist in finding how to take advantage of the exhaustive sampling made possible in simple lattice models to reveal other hidden, unknown, or poorly known properties of natural proteins.

Financial support was provided by the CNRS through the Genome Project. We thank IDRIS (Orsay, France) for a grant of computing time (Grant No. c980777).

[1] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).

[2] A. Sali, E. I. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).

[3] E. I. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).

[4] C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, edited by P. Debrunner, J. C. M. Tsibris, and E. Munch

(University of Illinois Press, Urbana, 1969), pp. 22–24.
[5] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[6] A. V. Finkelstein and O. B. Ptitsyn, Prog. Biophys. Mol. Biol. **50**, 171 (1987).
[7] C. Chothia, Nature (London) **357**, 543 (1992).
[8] D. Bashford, C. Chothia, and A. M. Lesk, J. Mol. Biol. **196**, 199 (1987).
[9] H. Li, C. Tang, and N. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[10] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[11] R. C. L. Milton, S. C. F. Milton, and S. B. H. Kent, Science **256**, 1445 (1992).
[12] M. Kimura, Nature (London) **217**, 624 (1968); *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
[13] T. P. Flores, C. A. Orengo, D. S. Moss, and J. M. Thornton, Protein Sci. **2**, 1811 (1993).
[14] S. Dalal, S. Balasubramanian, and L. Regan, Nat. Struct. Biol. **4**, 548 (1997).
[15] G. D. Rose and T. P. Creamer, Proteins: Struct., Funct., Genet. **19**, 1 (1994); G. D. Rose, Nat. Struct. Biol. **4**, 512 (1997).
[16] S. B. Prusiner, Arch. Neurol. Chicago **50**, 1129 (1993).
[17] The 27-mer chain cannot start or end at the cube center or at any edge center [18,19]. Interestingly, this eliminates here the spurious sticky-end effect.
[18] O. I. Razgulyaev, J. Chem. Phys. **93**, 5968 (1990).
[19] O. Bokanovsky (private communication).
[20] C. Branden and J. Toose, *Introduction to Protein Structure* (Garland, New York, 1991), pp. 12 and 38.
[21] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 638 (1990).
[22] The two remaining $5H$ sequences and the single $4H$ sequence all have a low gap of 0.3 and belong to the same neutral island.
[23] M. R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, Phys. Rev. E **57**, 3298 (1998).
[24] C. Tang and N. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998).
[25] As mentioned, this number ($N_i = 3790$) is closed to but not identical to the value of N_S found in Ref. [5] ($N_S = 3794$).
[26] M. Gerstein, A. M. Lesk, and C. Chothia, Biochemistry **33**, 6740 (1994).
[27] S. J. Gould and N. Eldredge, Paleobiology **3**, 115 (1977).