Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues

RUSSELL SCHWARTZ,[1] SORIN ISTRAIL,[2] AND JONATHAN KING[1]

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[2]Celera Genomics, Rockville, Maryland 20850, USA

(RECEIVED August 1, 2000; FINAL REVISION January 23, 2001; ACCEPTED February 27, 2001)

Abstract

Patterns of hydrophobic and hydrophilic residues play a major role in protein folding and function. Long, predominantly hydrophobic strings of 20–22 amino acids each are associated with transmembrane helices and have been used to identify such sequences. Much less attention has been paid to hydrophobic sequences within globular proteins. In prior work on computer simulations of the competition between on-pathway folding and off-pathway aggregate formation, we found that long sequences of consecutive hydrophobic residues promoted aggregation within the model, even controlling for overall hydrophobic content. We report here on an analysis of the frequencies of different lengths of contiguous blocks of hydrophobic residues in a database of amino acid sequences of proteins of known structure. Sequences of three or more consecutive hydrophobic residues are found to be significantly less common in actual globular proteins than would be predicted if residues were selected independently. The result may reflect selection against long blocks of hydrophobic residues within globular proteins relative to what would be expected if residue hydrophobicities were independent of those of nearby residues in the sequence.

Keywords: Sequence database; bioinformatics; aggregation; inclusion body; hydrophobic residues

The distribution of hydrophobic versus hydrophilic residues along polypeptide chains is a critical feature of the ability of biologically evolved amino acid sequences to direct the folding of proteins. Among well-documented features are the burial of hydrophobic residues to form the cores of globular proteins (Rose and Roy 1980), the alternation of hydrophobic and hydrophilic residues in β strands, the heptad repeat pattern in coiled-coil strands (Cohen and Parry 1986; Berger et al. 1995), and the very long strings of hydrophobic residues in transmembrane helices (Tomita and Marchesi 1975). The latter feature has proved very useful to biochemists and cell biologists for identifying putative transmembrane proteins in the absence of structural information (von Heijne 1994). In the course of investigating the sequence determinants of polypeptide chain interactions that compete with productive folding, we became interested in hydrophobic runs within globular proteins.

The failure of protein folding has emerged as an important biological and biotechnological problem (Mitraki and King 1989). A common failure mode is the self-association of folding intermediates, leading to an aggregated biological state (Speed et al. 1997; Wetzel 1997). In a number of well-studied systems this off-pathway reaction has been shown to be associated with partially folded intermediates. A key role of heat shock chaperonins in cells is recognizing junctional misfolded intermediates and helping them avoid self-association and the kinetically trapped aggregated state. The isolation of mutants influencing this partitioning establishes that it is a property of amino acid sequences (Mitraki et al. 1991).

In recent years, several increasingly detailed computational and theoretical models have been developed to ad-

Reprint requests to: Dr. Jonathan King, Department of Biology, Massachusetts Institute of Technology, Room 68-330, Cambridge, MA 02139, USA; e-mail: jaking@mit.edu; fax: (617) 252-1843.

Article and publication are at www.proteinscience.org/cgi/doi/10.1110/ps.33201.

dress questions of aggregate structure and dynamics currently inaccessible to experimental techniques (De Young et al. 1993; Patro and Przybycien 1994, 1996; Gupta et al. 1998). In prior work, we reported on a lattice-based computer simulation model for studying the propensity of polypeptide chains to aggregate during simulated folding in a solution (Istrail et al. 1999). The key idea of the method was to follow two identical chains folding independently and periodically test how the folding intermediates could associate with one another in an energetically favorable manner. On the basis of our results, we concluded that propensity of a sequence to aggregate within the model is a consistent property of its amino acid sequence and correlates with measurable properties of the sequence.

One particular property found to predict a propensity to aggregate was the grouping of hydrophobic residues within a peptide into small numbers of consecutive blocks. For a given number of hydrophobic residues, the computer model showed increasing propensity to aggregate as the hydrophobic residues became concentrated in fewer continuous subsequences. This property is of interest for the present work because it implies a testable prediction about actual sequences known to fold in aqueous solution: they will have evolved to select against long blocks of consecutive hydrophobic residues in order to promote a low loss of proteins to aggregation. The present work examines whether selection against long hydrophobic sequences within globular proteins is reflected in actual published sequence data. Our methodology is therefore similar to that adopted by Broome and Hecht (2000), who studied statistical distributions of small patterns of hydrophobic and hydrophilic residues to support the hypothesis that a pattern of alternating hydrophobic–hydrophilic residues predisposes sequences to aggregation.

The specific issue of hydrophobic run lengths addressed by the present work has been previously examined by White and Jacobs (1990). They performed a careful statistical analysis of extant protein sequences, testing the probability—given the hypothesis that residue hydrophobicities are assigned independently at random—of the number of hydrophobic or hydrophilic runs differing from the expectation by at least as much as that observed in each individual sequence. They concluded that the majority of individual proteins examined have distributions of hydrophobic run lengths statistically indistinguishable from those expected for random sequences. White and Jacobs (1990) argued from this result in favor of the hypothesis that extant protein sequences are essentially random except at small numbers of conserved sites, consistent with the evolutionary hypothesis that current biologically relevant proteins evolved from a large initial pool of random sequences.

Although the result of White and Jacobs (1990) appears to argue against our hypothesis, the questions we ask differ from those originally examined by them, primarily in that we limit our search to proteins known to fold in aqueous solution and in that we compile statistics only for an entire database, rather than for individual sequences. This latter difference is subtle but important in allowing genuine statistical effects that are small in absolute magnitude to rise to the level of statistical significance. In later work, White and Jacobs (1993) explored aggregate statistics on run lengths of hydrophobic versus hydrophilic residues and found a slight bias toward shorter consecutive blocks. We extend these studies primarily in examining statistics for proteins known to fold in aqueous solution and in contrasting those results to statistics derived from known membrane sequences and to complete proteomes.

Polypeptide chains that do refold to their native state in aqueous buffers, generally fail to reach this state in the presence of detergents, lipid vesicles, or organic solvents. In contrast, the refolding of integral membrane proteins in vitro requires both lipid vesicles and surfactants. These experimental observations together with theoretical considerations indicate that the folding of these two classes of proteins proceeds through very different intermediates and pathways, and requires different environmental conditions.

The overall statistical content of protein sequences has since been examined by Strait and Dewey (1996), who analyzed a protein sequence database in terms of its information entropy by different measures of information content. They determined that actual protein sequences carry significantly less information than is theoretically possible from a 20-letter alphabet. Analyses specifically focusing on hydrophobicity have mainly been aimed at screening for transmembrane segments of proteins via hydrophobicity scales, a technique pioneered by Kyte and Doolittle (1980). Subsequent experience has borne out the association of long hydrophobic stretches with transmembrane helices.

Results and Discussion

For the present work, sequences were initially taken from the ASTRAL compendium of protein sequences (Brenner et al. 2000) based on version 1.48 of the SCOP structural classification of sequences (Murzin et al. 1995). The sequence file was produced by an ASTRAL request for sequences having no more than 50% sequence identity in order to yield a database eliminating problems of redundancy. The database was then screened to remove those assigned to SCOP class six, membrane and cell-surface proteins and peptides. Finally, sequences were excluded from the statistical analysis if they contained any amino acids of unspecified type. The resulting database contained 472,286 amino acids in 2753 sequences.

A quantitative measure of the concentration of hydrophobic residues into blocks is the number of alternations in the database, in which a hydrophobic residue is immediately followed by a nonhydrophobic residue or vice versa; the

sum of this quantity over all 2753 sequences was computed. In addition, a histogram was recorded of the lengths of all maximal sequences of consecutive hydrophobic residues within the protein sequences analyzed. Hydrophobic residues were defined for the purposes of this study to be Ala, Ile, Leu, Met, Phe, Pro, Trp, Tyr, and Val.

Data were also recorded on the distribution of sequence lengths in the database and on the probability p_H that a randomly selected residue within the database was hydrophobic. A prediction of the expected number of alternations given the frequency of hydrophobic residues and the sequence lengths in the database was calculated. Predicted values were also derived computationally for the expected number of blocks of consecutive hydrophobic residues of each possible length, given the measured distribution of sequence lengths, assuming that each residue was assigned independently at random to be hydrophobic with probability p_H, as described in the Materials and Methods section. The result provides a measure of the expected block length frequencies given a sufficiently large database and the assumption that no factors influence the relative positions of hydrophobic residues within each sequence.

Both measures suggest long blocks of hydrophobic residues are suppressed relative to what would be expected if residues were chosen independently of their neighbors. For our definition of hydrophobic residues, p_H was found to be 0.451. The expected number of alternations given this value of p_H was calculated as 232,542, compared to a measured value of 237,716. The difference of 5,174 (2.225% of expected) is 14.97 standard deviations and is therefore unlikely to be caused by chance. Figure 1 shows the measured and expected values of hydrophobic block lengths for non-membrane-associated proteins. The measured values slightly exceed those predicted under the assumption of independence of residues for hydrophobic block lengths one, two, and three. However, they fall significantly lower for all longer block lengths except 16, for which one block was detected whereas 0.388607 blocks were expected. No hydrophobic block lengths longer than 16 were observed in the sequence data. Table 1 lists those sequences containing hydrophobic blocks of length 12 or longer, identified by PDB ID (Bernstein et al. 1977).

Can the above results be accounted for by a relative increase in certain motifs containing short hydrophobic blocks, such as the alternating hydrophobic–hydrophilic pattern characteristic of surface β sheets? Any relative decrease in hydrophobic residues in long blocks relative to expectations must be exactly offset by a corresponding increase in hydrophobic residues in other block lengths above that expected. The surplus hydrophobic residues are found at block lengths 1, 2, 3, and 16, with a total of 700 surplus hydrophobic residues in length 1 blocks (9.81% of the total surplus), 5478 surplus hydrophobic residues in length 2 blocks (76.8%), 945 surplus hydrophobic residues in length 3 blocks (13.2%), and 9.78 surplus hydrophobic residues in a single length 16 block (0.137%).

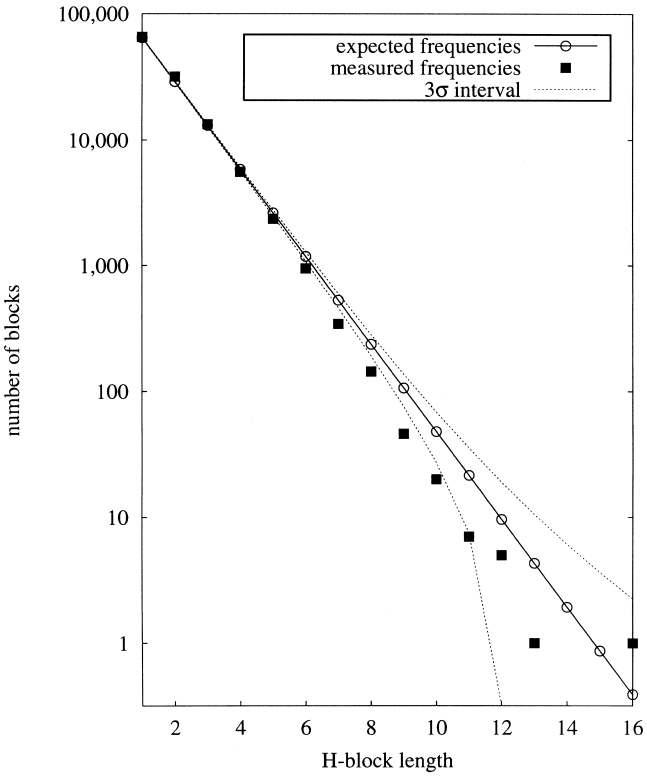

Fig. 1. Semi-log plot of measured values of hydrophobic block length distributions and expected values for non-membrane-associated proteins given the assumption of independence between residues. No measured value is shown for lengths 14 or 15 because no blocks of those lengths were found. Dashed lines show an estimated region of plus or minus three standard deviations about the expectation.

How important is our specific choice of residues to classify as hydrophobic? To examine this question, we repeated our statistical calculations on numbers of alternations for the database for other possible selections of hydrophobic residues. We altered our set of hydrophobic residues in 20 successive recalculations by individually subtracting out each amino acid we classified as hydrophobic and by individually adding in each residue we classified as nonhydrophobic. The resulting data are summarized in Table 2. In each case, a single amino acid change still results in a statistically significant elevation in the number of alternations, although usually to a lesser degree than with our originally chosen set of hydrophobics. The most dramatic increase occurs with the addition of glycine, increasing the percent elevation over the expectation from 2.225% to 2.964%. The most dramatic decrease comes from adding in aspartate, decreasing the percent elevation to 0.968%.

Because all 21 sets examined have elevated alternations relative to their expected values, it might be conjectured that all or almost all possible choices of hydrophobic residues

Table 1. *Sources of observed hydrophobic blocks of length 12 or greater from the database of non-membrane-associated sequences*

PDB ID	Chain or fragment	Protein description	Longest continuous hydrophobic block
1TCA		Triacylglycerol hydrolase *(Candida antarctica)*	16 in surface helix and bend
1GUQ	A2	Galactose-1-phosphate uridylyltransferase mutant *H166G (Escherichia coli)*	13 in buried strand and turn
1ALN	1	Cytidine deaminase *(E. coli)*	12 in surface helix
2MAD	H	Methylamine dehydrogenase *(Thiobacillus versutus)*	12 in solvent-exposed helix
2PHL	A2	Phaseolin *(Phaseolus vulgaris)*	12 in half-buried strand–turn–strand
1UAE		UDP-*N*-acetylglucosamine enolpyruvyl transferase *(E. coli)*	12 in buried helix and turn
1UBY		Farnesyl diphosphate synthase mutant *F112A, F113S (Gallus gallus)*	12 in buried helix

This table is not necessarily unique because homologous sequences were screened from the database.

Table 2. *Alternations for different sets of "hydrophobic" amino acids*

Set[a]	Obs.[b]	Exp.[c]	Δ/σ[d]	$\%\Delta$[e]
Σ	237716	232542	14.96	2.225
Σ – Ala	224470	218844	15.51	2.571
Σ – Phe	230524	227455	8.704	1.349
Σ – Ile	227087	224572	7.061	1.120
Σ – Leu	220239	217731	6.891	1.152
Σ – Met	235169	230113	14.48	2.197
Σ – Pro	230772	226207	12.89	2.018
Σ – Val	225619	221585	11.21	1.821
Σ – Trp	235153	231052	11.79	1.775
Σ – Tyr	231747	228012	10.61	1.638
Σ + Cys	239792	233818	17.37	2.555
Σ + Asp	236935	234663	6.628	0.968
Σ + Glu	237440	234518	8.520	1.246
Σ + Gly	241054	234114	20.20	2.964
Σ + His	238599	234145	12.97	1.902
Σ + Lys	238622	234646	11.60	1.694
Σ + Asn	238494	234756	10.91	1.592
Σ + Gln	238502	234663	11.20	1.636
Σ + Arg	238207	234766	10.04	1.466
Σ + Ser	238702	234630	11.88	1.735
Σ + Thr	237489	234702	8.132	1.187

[a] The set of residues counted as hydrophobic relative to our base set Σ.
[b] The observed number of alternations between hydrophobic and hydrophilic.
[c] The expected number given the assumption of independence between residue positions.
[d] The difference between the expected and observed values in standard deviations.
[e] The distance between expected and observed values as a percentage of the expected value.

would yield elevated values, regardless of whether the choice actually reflects a biologically meaningful grouping of residues by hydrophobicity. To assess that possibility, we calculated the percent elevation over expectation of alternations for our database for all possible partitions of the 20 amino acids into two nonempty groups. For the 524,287 possible groups (2^{20}, minus the empty and full sets, divided by 2 because of the symmetry of the calculations with respect to which set is called hydrophobic), we found a mean decrease in alternations relative to expectation of 0.358%, with a standard deviation of 0.593%. Among individual groupings, 399,522 showed decreased alternations (211,647 by more than 3 standard deviations), whereas 124,765 showed increased alternations (37,721 by more than 3 standard deviations). This result suggests that the significantly elevated numbers of alternations observed for our original assignment of hydrophobic residues and for the single amino acid variants of it are atypical of most groupings, which are shifted noticeably toward decreased alternations. The results therefore appear to depend on the fact that they are derived from a biologically reasonable choice of hydrophobic residues, but they do not critically depend on the exact residue composition of that group.

Does removing membrane and cell-surface proteins from the database introduce a sample bias? This decision should not undermine the detection of a pattern in the database if it reflects properties of polypeptide chains from globular proteins that fold in aqueous solution. The selection of sequences to exclude from the analysis was based on this functional criterion, rather than on sequence. Furthermore, the expected values were computed using amino acid frequencies taken from the database after membrane and cell-surface proteins had been excluded, and the statistical suppression is therefore genuine given the overall hydrophobic content of non-membrane-associated sequences.

When similar statistics are computed from the excluded membrane and cell-surface proteins, also removing redundant sequences and those with amino acids of unspecified type, the results are qualitatively reversed. The histogram of expected and measured block frequencies for the membrane and cell-surface proteins is illustrated in Figure 2. The figure shows that long hydrophobic block frequencies are generally elevated relative to statistically expected values, assuming independent residue selection. However, the measured number of alternations is still slightly higher than expected with 6242 counted and 6144 expected, a difference of 1.595% of expected, or 1.764 standard deviations. Although we attribute the elevated frequencies largely to the contribution of transmembrane helices, blocks of the length required for a full transmembrane helix are not observed owing to the presence of nonhydrophobic residues within

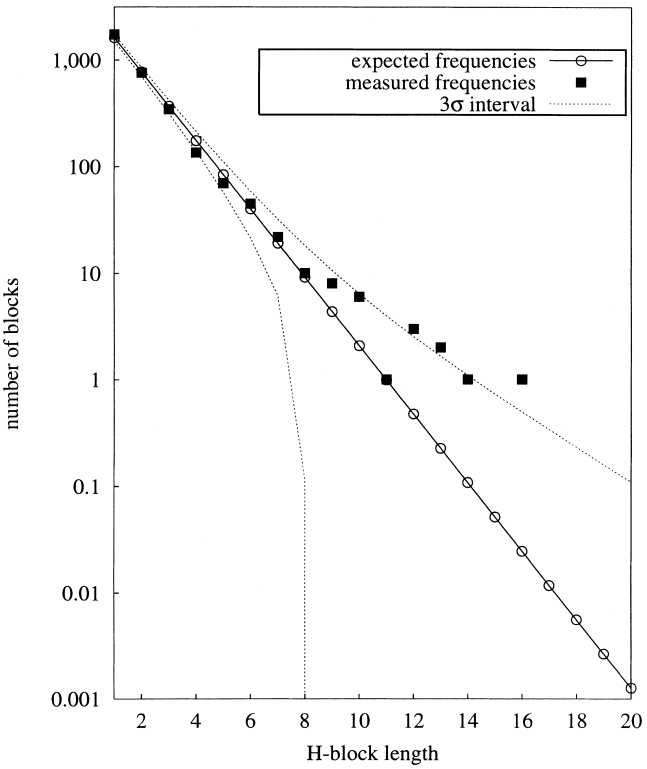

Fig. 2. Semi-log plot of measured and expected values of hydrophobic block length distributions for membrane and cell-surface proteins and peptides. No measured value is shown for length 15 or for any lengths greater than 16 because no blocks of those lengths were found. Dashed lines show an estimated region of plus or minus three standard deviations about the expectation.

the helices. Although our simulation model says nothing about membrane-associated sequences, the statistical data suggest that some of the effects predicted for sequences folding in aqueous solution are reversed for those that are membrane-associated.

A final question is whether the results for soluble proteins are typical of proteins in general. Statistics were compiled for complete databases of translated annotated open reading frames (ORFs) for *Escherichia coli* K12 (Blattner et al. 1997), *Saccharomyces cerevisiae* (Goffeau et al. 1996), and *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998). The *E. coli* data were produced by the *E. coli* Genome Project at the University of Wisconsin–Madison and can be obtained from ftp://ftp.genome.wisc.edu/pub/sequence/m52p.fap.gz. The *S. cerevisiae* data were produced by the Yeast Sequencing Group at the Sanger Centre and can be obtained from ftp://ftp.ebi.ac.uk/pub/databases/yeast/ORF_SEQS/protseq.pir. The *C. elegans* data were produced by the Worm Sequencing Group at the Sanger Centre and can be obtained from ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wp.fasta19. The databases varied in size and hydrophobicity, with 1,359,208 residues processed (48.7% hydrophobic) for *E. coli*, 2,981,004 residues (42.5% hydrophobic) for *S. cerevisiae*, and 10,166,480 residues (44.0% hydrophobic) for *C. elegans*. Figure 3 shows the expected and measured block frequencies for the three databases. All three are qualitatively similar, with elevated values of long block lengths and suppressed values for short block lengths, with the exception of an overabundance of blocks of length 2 in all three databases. The transition from underrepresented to overrepresented occurs at length 6 for *C. elegans*, 7 for *S. cerevisiae*, and 8 for *E. coli*. All three exhibit significantly fewer alternations between hydrophobic and hydrophilic than would be expected by chance: 674,843 counted versus 677,030 expected for *E. coli* (a difference of 0.3230%, or 3.755 σ), 1,431,462 counted versus 1,454,311 expected for *S. cerevisiae* (a difference of 1.571%, or 25.95 σ), and 4,885,131 counted versus 4,998,382 expected for *C. elegans* (a difference of 2.266%, or 70.13 σ).

The frequency analysis demonstrates that the database examined is missing sequences containing long strings of hydrophobic residues that would be expected to have occurred by chance if the distribution of hydrophobic residues within a sequence were unimportant. It is possible that this result reflects a sample bias arising from the use of a database of sequences whose structures have been solved or from some unanticipated evolutionary pressure.

It might be argued that fully folded proteins cannot tolerate long blocks of hydrophobic residues and remain soluble. However, we know of no reason why long hydrophobic blocks could not be accommodated internally to the structure of a globular protein after it has completed the folding process. Such structures can occur, as some of those listed in Table 2 illustrate. Figure 4 shows the structure of UDP-*N*-acetylglucosamine enolpyruvyl transferase, which accommodates its 12-residue hydrophobic block in a buried alpha helix, providing an excellent example of how a folded protein can accommodate a long string of consecutive hydrophobic residues. If such structures can be stable in a fully folded protein, then the question remains why they are statistically rare. We suggest that constraints imposed by the process of folding, as opposed to the structural needs of a fully folded protein, may be partially responsible for the observed effects. The results might also be an artifact of an overabundance of certain sequence motifs or elements of secondary structure that favor short hydrophobic blocks. The fact that the excess of short blocks consists primarily of length 2 blocks is not, however, consistent with any common structural motif known to us.

Although our primary result, that the set of protein sequences examined differs in hydrophobic block length distributions from what would be expected for a database of random strings, may initially appear to contradict that of White and Jacobs (1990), that most protein sequences do not differ significantly from random strings, the two results are not inconsistent. Part of the difference is likely ac-

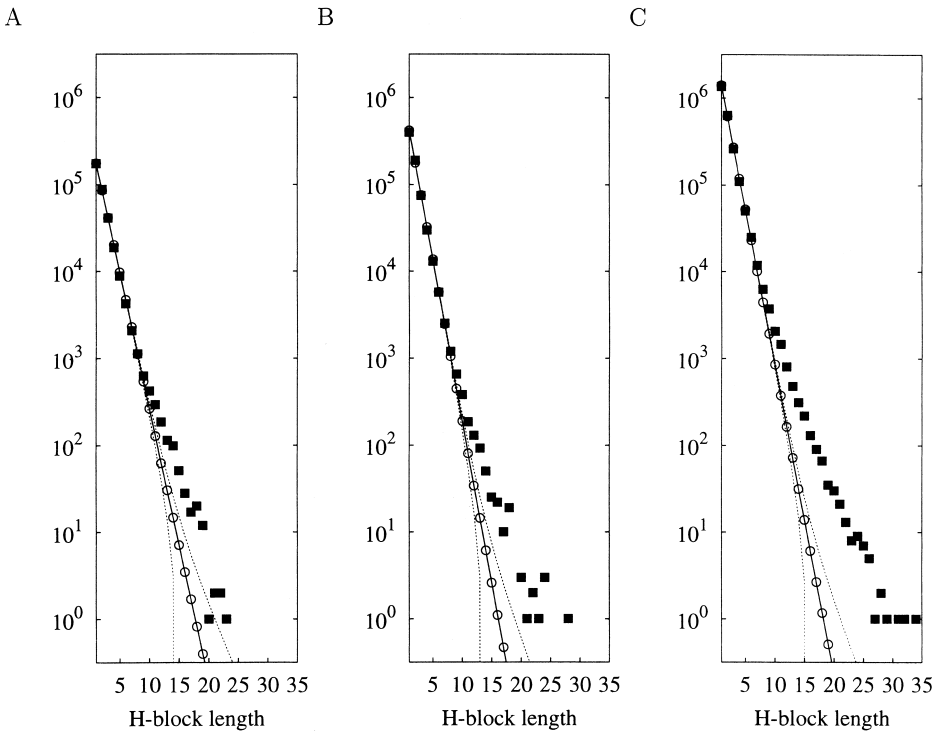

Fig. 3. Semi-log plots of measured and expected values of hydrophobic block length distributions for databases compiled from all open reading frames for a single organism. (*A*) *Escherichia coli* K12. (*B*) *Saccharomyces cerevisiae*. (*C*) *Caenorhabditis elegans*. Solid boxes show measured data points, circles calculated expected values, and dashed lines an estimated region of plus or minus three standard deviations about the expectation.

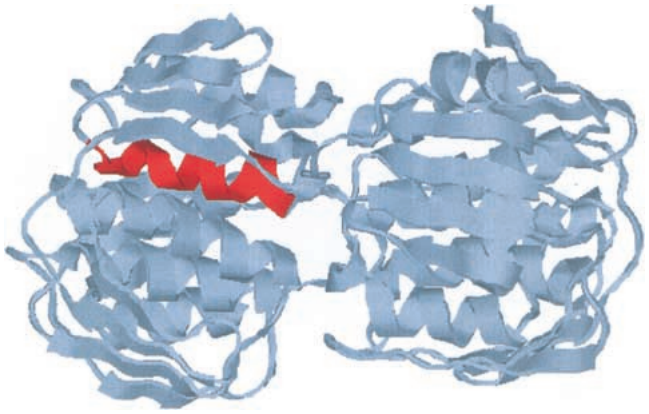

Fig. 4. A ribbon diagram of the structure of UDP-*N*-acetylglucosamine enolpyruvyl transferase as reported by Skarzynski et al. (1996). The 12 consecutive hydrophobic residues at positions 24–35 are highlighted. The image was prepared with RasMol version 2.6.

counted for by the differences in construction of a dataset for analysis, as we limited ourselves to cytosolic proteins of known structure. Even beyond that, our methodology differs somewhat from that originally used by White and Jacobs (1990) in that we examine statistics only for the database as a whole, rather than asking questions about isolated sequences. The fact that the statistical disparities found in frequencies of long block lengths and alternations rise to the level of statistical significance over the database as a whole does not imply that more than a small fraction of individual sequences are statistically distinguishable from random sequences.

The disparity in alternations is small as a percentage of total alternations and would not be expected to reach statistical significance in individual sequences if it were distributed evenly across the database, even though it is quite significant for the database as a whole. Similarly, although disparities in block lengths show up as significant in the database as a whole, the absolute number of residues expected to be involved in long blocks is a small percentage of the total number of residues. A disparity in those frequencies over the whole proteome may therefore not appear to be statistically significant when individual sequences are examined in isolation. We believe that our data and those of White and Jacobs (1990, 1993) are both consistent with two interpretations: a small systematic bias in block lengths across the whole database or large differences in a minority of sequences and no differences in others. We are not aware of a method for deciding between these two hypotheses. We therefore believe the apparent disparity between our results and those of White and Jacobs (1990, 1993) reflects our

asking subtly different questions, guided by a concern for the influence of sequence on the process of folding as opposed to final folded states alone. Neither do our results rule out the premise of White and Jacobs that sequences of biologically relevant proteins may have evolved from an initially random set, with the caveat that selective pressures may have eliminated a noticeable subset of that initial set.

On the basis of our prior simulation results, we suggest that the missing sequences reflect in part the evolutionary fitness constraint imposed by selective pressure to avoid off-pathway aggregation. As proteins evolved to favor sequences that could fold reliably, a noticeable suppression of sequences with long strings of consecutive hydrophobic residues occurred. This effect could create a counterbalance to the drive for high hydrophobicity identified by Moult and Unger (1991) as a key predictor of rapid folding, creating the need for a balance between these two competing factors of folding rate and aggregation propensity. This conclusion is consistent with results from our earlier lattice simulations (Istrail et al. 1999), which found that rapid folders in our lattice model tended to have high aggregabilities per unit time, apparently because of a correlation of both fast folding and high propensity to aggregate with high hydrophobic content. The reversed results seen for membrane-associated proteins may reflect how evolutionary selection for sequence operates differently on membrane-associated sequences than on those evolved for aqueous environments. In itself, the similarity between the membrane-associated sequences and the complete-ORF databases does not necessarily mean that membrane-associated proteins form a large fraction of actual ORFs, only that complete genomes contain many membrane-associated proteins that are able to produce a significant effect on the data at long block lengths, where soluble proteins produce few data points. However, the fact that elevations of long blocks in complete genomes are more pronounced than in the membrane proteins of solved structure does suggest that the solved membrane proteins are not fully representative of membrane proteins in general.

The results from membrane-associated and complete-ORF databases strengthen our contention that some constraint imposed by the requirements of folding in aqueous solution is suppressing the number of hydrophobic residues found in long consecutive blocks in soluble proteins. These conclusions suggest that aggregation constraints may contribute to the observation of Strait and Dewey (1996) that actual protein sequences carry considerably less information than a 20-letter alphabet theoretically allows. They further suggest the importance of considering propensity to aggregate as a design constraint in protein evolution on a par with rapid folding and with stability and functional fitness of the native state. The confirmation of an important statistical prediction of our abstract computer model also supports the validity of that model and suggests the benefits of refining computational methodologies for exploring protein folding and aggregation. The tremendous increase in data available for analysis in recent years suggests that similar database analysis techniques for locating interesting proteins or supporting general hypotheses about protein behavior are likely to become increasingly valuable.

Materials and methods

Expected numbers of alternations for a single sequence are given in terms of the sequence length n and the probability p_H that a residue chosen at random is hydrophobic, by $2(n-1)p_H(1-p_H)$. Summing this quantity over all sequence lengths in the database gives the total expected number of alternations. The variance of the distribution given the assumption of independence between residues can be calculated for a single sequence of length n by summing the variances of all pairs of consecutive residues, which is $(n-1)[2p_H(1-p_H) - 4(p_H)^2(1-p_H)^2]$, and adding twice the sum of the covariances between overlapping pairs of consecutive residues, which is $2(n-2)[p_H(1-p_H) - 4(p_H)^2(1-p_H)^2]$. The sum of these quantities over all sequence lengths is the total expected variance for the database, and the standard deviation is the square root of this variance.

Expected hydrophobic block length frequencies given the assumption of independence between positions were calculated via a function $P(n,k,m)$, expressing the probability for fixed p_H that a sequence of length n has exactly m hydrophobic blocks of length exactly k. $P(n,k,m)$ can be calculated via a recurrence relation by considering the following cases:

m = *0,* n < k

There can be no length k hydrophobic blocks in a string of length n for $n < k$, therefore $P(n,k,0) = 1$.

m = *0,* n = k

If $n = k$, then either there are no length k hydrophobic blocks or the string is entirely hydrophobic, with probability $(p_H)^k$, in which case there is exactly 1. Therefore $P(n,k,0) = 1 - (p_H)^k$.

m = *0,* n > k

If $n > k$, m can be zero only if the string is entirely hydrophobic, with probability $(p_H)^n$, or for some $i \neq k$ the string consists of i hydrophobic residues, followed by a hydrophilic residue, followed by a string with no length k hydrophobic blocks. Therefore,

$$P(n,k,0) = (p_H)^n + \sum_{i=0}^{k-1} (p_H)^i(1-p_H)P(n-i-1,k,0) + \sum_{i=k+1}^{n-1} (p_H)^i(1-p_H)P(n-i-1,k,0).$$

m > *(*n + *1)/(*k + *1)*

In order to satisfy $m > 0$, there must be mk hydrophobic residues in the m hydrophobic blocks and a minimum of $(m-1)$ hydrophilic residues separating the blocks. This requirement cannot be satisfied if $m > (n+1)/(k+1)$, therefore $P(n,k,m) = 0$.

m = *(*n + *1)/(*k + *1)*

There is exactly one arrangement of residues that gives m hydrophobic blocks of length k, in which each pair of consecutive blocks is separated by one hydrophilic residue. The probability of this single arrangement is the probability of choosing the mk hydrophobic residues and the $(m-1)$ hydrophilic residues, giving $P(m,n,k) = (p_H)^{mk}(1-p_H)^{m-1}$.

m = *1,* n > k

This relationship can be satisfied only if the string begins with a k-block, followed by a hydrophilic residue, followed by a string containing no k-blocks; it ends with a k-block, preceded by a hydrophilic residue, preceded by a string with no k-blocks; or for some i it consists of a k-block surrounded by two hydrophilic residues, surrounded by two blocks, one of length $(i-1)$, which contain no k-blocks. Therefore,

$$(p_H)^k(1-p_H)P(n-k-1,k,0) + P(n-k-1,k,0)(1-p_H)(p_H)^k + \sum_{i=1}^{n-k-1} P(i-1,k,0)(1-p_H)(p_H)^k(1-p_H)P(n-i-k-1,k,0).$$

otherwise

This case captures the condition that $m > 1$ and the string is long enough to contain m k-blocks. The probability for this case can be specified by considering the location of the first k-block. Either the string begins with a k-block, followed by a hydrophilic residue, followed by a string containing $(m-1)$ k-blocks, or, for some i, it begins with a string of length $(i-1)$ containing no k-blocks, followed by the first k-block surrounded by two hydrophilic residues, followed by a string containing $(m-1)$ k-blocks. This gives:

$$P(n,k,m) = (p_H)^k(1-p_H)P(n-k-1,k,m-1) + \sum_{i=1}^{n-k-1} P(i-1,k,0)(1-p_H)(p_H)^k(1-p_H) P(n-i-k-1,k,m-1).$$

Given $P(n,k,m)$, the expected number of hydrophobic blocks of length k in a single sequence of length n is given by:

$$H(n,k) = \sum_{i=0}^{\lfloor n/k \rfloor} iP(n,k,i).$$

The sum over all sequence lengths n of $H(n,p_H,k)$ multiplied by the multiplicity of sequences of length n gives the overall estimated frequency of blocks of length k for each k, given the assumption that amino acid types are mutually independent. The recurrence relation derived from the above analysis, with some mathematical simplifications, was used in the present work to calculated expected block length distributions.

Standard deviations of the expected block counts were estimated via simulations. Simulation trials were performed by creating a random database of sequences with the same lengths as those in the measured database but with residues assigned independently at random, with the measured probability p_H that a given residue was hydrophobic. Block lengths within the random database were then measured as was done with the actual data. Standard deviations were computed based on 1,000,000 such trials.

When computing statistics for all possible partitions of the set of amino acids into two groups, we first counted the number of occurrences of all amino acids and all pairs of consecutive amino acids in the databases. We calculated the number of alternations for each partition by summing the counts of all consecutive residue pairs assigned to different groups by the partition. We similarly calculated p_H by dividing the sum of the values for residues classified as hydrophobic in a given partition by the total number of amino acids in the database. Given p_H, we then calculated the expectation and standard deviation of numbers of alternations for each partition as described above.

Acknowledgments

This work was supported in part by Sandia National Laboratories, operated by Lockheed Martin for the U.S. Department of Energy under contract No. DE-AC04-94AL85000 and by the Mathematics, Information, and Computational Science Program of the Office of Science of the U.S. Department of Energy. It was also supported by the U.S. National Institutes of Health under grant NIH GM 17,980. We thank Lenore Cowen, Phil Bradley, and Peter Thumfort for their criticisms of our statistical arguments and suggesting means to strengthen them.

The publication costs of this article were defrayed in part by payment of page charges.This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Berger, B., Wilson, D.B., Wolf, E., Tonchev, T., Milla, M., and Kim, P.S. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci. USA* **92:** 8259–8263.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112:** 535–542.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462.

Brenner, S.E., Koehl, P., and Levitt, M. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucl. Acids Res.* **28:** 254–256.

Broome, B.M. and Hecht, M.H. 2000. Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis. *J. Mol. Biol.* **296:** 961–968.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Cohen, C. and Parry, D.A.D. 1986. α-Helical coiled coils—A widespread motif in proteins. *Trends Biochem. Sci.* **11:** 245–248.

De Young, L.R., Fink, A.L., and Dill, K.A. 1993. Aggregation of globular proteins. *Accounts Chem. Res.* **26:** 614–620.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546–567.

Gupta, P., Hall, C.K., and Voegler, A.C. 1998. Effect of denaturant and protein concentrations upon protein refolding and aggregation: A simple lattice model. *Protein Sci.* **7:** 2642–2652.

Istrail, S., Schwartz, R., and King, J.A. 1999. Lattice simulations of aggregation funnels for protein folding. *J. Comp. Biol.* **6:** 143–162.

Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105–132.

Mitraki, A. and King, J. 1989. Protein folding intermediates and inclusion body formation. *Bio/Technology* **7:** 690–697.

Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J., and King, J. 1991. Global suppression of protein folding defects and inclusion body formation. *Science* **253:** 54–58.

Moult, J. and Unger, R. 1991. An analysis of protein folding pathways. *Biochem.* **30:** 3816–3824.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A

structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Patro, S.Y. and Przybycien, T.M. 1994. Simulations of kinetically irreversible protein aggregate structure. *Biophys. J.* **66:** 1274–1289.

———. 1996. Simulations of reversible protein aggregate and crystal structure. *Biophys. J.* **70:** 2888–2902.

Rose, G.D. and Roy, S. 1980. Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA* **77:** 4643–4647.

Skarzynski, T., Mistry, A., Wonacott, A., Hutchinson, S.E., Kelly, V.A., and Duncan, K. 1996. Structure of UDP-*N*-acetylglucosamine enolpyruvyl transferase, an enzyme essential for the synthesis of bacterial peptidoglycan, complexed with substrate UDP-*N*-acetylglucosamine and the drug fosfomycin. *Structure* **4:** 1465–1474.

Speed, M.A., King, J., and Wang, D.I.C. 1997. Polymerization mechanism of polypeptide chain aggregation. *Biotech. Bioeng.* **54:** 333–343.

Strait, B.J. and Dewey, T.G. 1996. The Shannon information entropy of protein sequences. *Biophys. J.* **71:** 148–155.

Tomita, M. and Marchesi, V.T. 1975. Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proc. Natl. Acad. Sci. USA* **72:** 2964–2968.

von Heijne, G. 1994. Membrane proteins: From sequence to structure. *Ann. Rev. Biophys. Biomol. Struct.* **23:** 167–192.

Wetzel, R. 1997. Protein misassembly. *Adv. Prot. Chem.* **50:** 330–350.

White, S.H. and Jacobs, R.E. 1990. Statistical distribution of hydrophobic residues along the length of protein chains: Implications for protein folding and evolution. *Biophys. J.* **57:** 911–921.

———. 1993. The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.* **36:** 79–95.