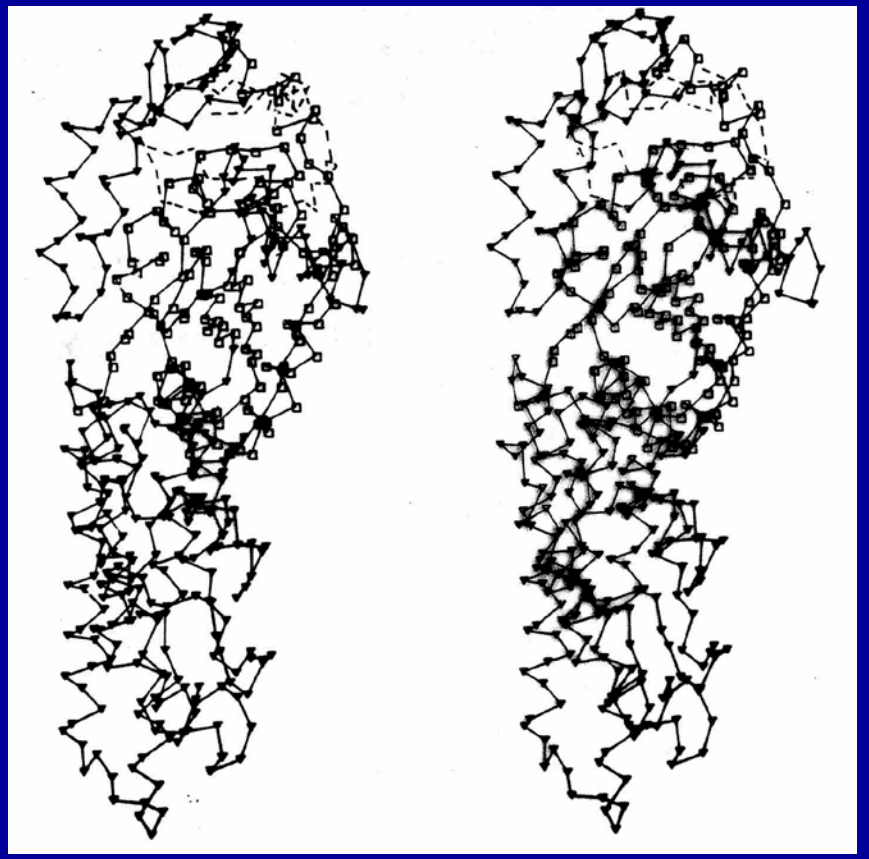
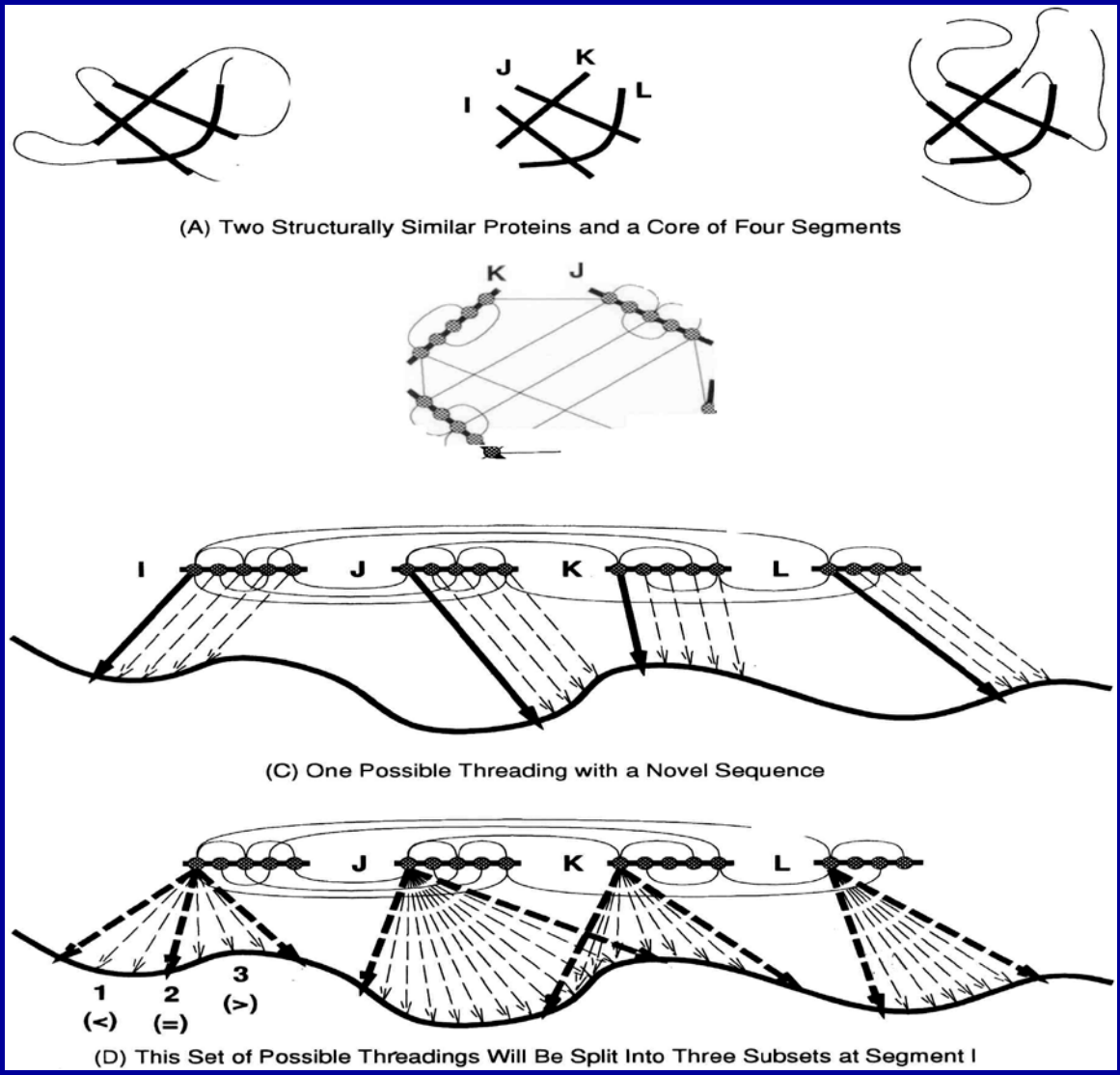
Protein structure prediction

TQVAKKILVTCALPYANGSIHLGHMLEHIQADVWVRYQRMRG
HEVNFICADDAHGTPIMLKAQQLGITPEQMIGEMSQEHQTDF
AGFNISYDNYHSTHSEENRQLSELIYSRLKENGFIKNRTISQLY
DPEKGMFLPDRFVKGTCPKCKSPDQYGDNCEVCGATYSPTEL
IEPKSVVSGATPVMRDSEHFFFDLPSFSEMLQAWTRSGALQEQ
VANKMQEWFESGLQQWDISRDAPYFGFEIPNAPGKYFYVWLD
APIGYMGSFKNLCDKRGDSVSFDEYWKKDSTAELYHFIGKDI
VYFHSLFWPAMLEGSNFRKPSNLFVHGYVTVNGAKMSKSRGT
FIKASTWLNHFDADSLRYYYTAKLSSRIDDIDLNLEDFVQRVN
ADIVNKVVNLASRNAGFINKRFDGVLASELADPQLYKRFTDA
AEVIGEAWESREFGKAVREIMALADLANRYVDEQAPWVVAK
QEGRDADLQAIAQWGINLFRVLMTYLKPVLPKLTERAEAFLN
TELTWDGIQQPLLGHKVNPFKALYNRIDMRQVEALVEASKEE
VKAAAAPVTGPLADDPQDGCGRHDRVVDSGSK



“Protein Threading”



Protein Engineering vol.7 no.9 pp. 1059–1068. 1994

The protein threading problem with sequence amino acid interaction preferences is NP-complete

Richard H.Lathrop

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

In recent protein structure prediction research there has been a great deal of interest in using amino acid interaction preferences (e.g. contact potentials or potentials of mean force) to align ('thread') a protein sequence to a known structural motif. An important open question is whether a polynomial time algorithm for finding the globally optimal threading is possible. We identify the two critical conditions governing this question: (i) variable-length gaps are admitted into the alignment, and (ii) interactions between amino acids from the sequence are admitted into the score function. We prove that if both these conditions are allowed then the protein threading decision problem (does there exist a threading with a score $\leq K$?) is NP-complete (in the strong sense, i.e. is not merely a number problem) and the related problem of finding the globally optimal protein threading is NP-hard. Therefore, no polynomial time algorithm is possible (unless P = NP). This result augments existing proofs that the direct protein folding problem is NP-complete by providing the corresponding proof for the 'inverse' protein folding problem. It provides a theoretical basis for understanding algorithms currently in use and indicates that computational strategies from other NP-complete problems may be useful for predictive algorithms.

Key words: contact potentials/inverse protein folding/NP-complete/protein structure prediction/protein threading/sequence–structure alignment

tion (e.g. secondary structure) constraints (Ngo and Marks, 1992; Fraenkel, 1993; Unger and Moult, 1993).

One important alternative approach is to use the known protein structures as (i) spatial folding templates, (ii) additional knowledge about protein structure and (iii) constraints on possible folds. This is a powerful strategy because folded proteins exhibit recurring patterns of organization. Chothia (1992) estimates that there are only ~1000 different protein structural families. In this approach, each known protein structure (or family) 'recognizes' the protein sequences likely to fold into a similar structure. Because it starts with structures and predicts sequences instead of starting with sequences and predicting structures, it is often referred to as the 'inverse' folding problem. In its fully general sense it includes *ab initio* design of protein sequences to achieve a target structure (Pabo, 1983), but we shall restrict attention to folding given native sequences. The known structure establishes a set of possible amino acid positions in 3-D space (perhaps the spatial locations of its main-chain α carbons). 'Recognition' is mediated by a suitable score function. An alignment between spatial positions and sequence amino acids is usually a by-product of the recognition step. The sequence is given a similar 3-D fold by placing its amino acids into their aligned spatial positions. [Further techniques are necessary to correctly place the variable loop regions (Greer, 1990; Zheng *et al.*, 1993) and position the side chains (Desmet *et al.*, 1992), but the focus of this paper is on predicting and placing the conserved fold.] The process of aligning a sequence to a structure and thereby guiding the spatial placement of sequence amino acids is referred to as 'threading' the sequence onto the structure (Bryant and Lawrence, 1993). 'A threading' means a specific

Heuristic (Objective) Function

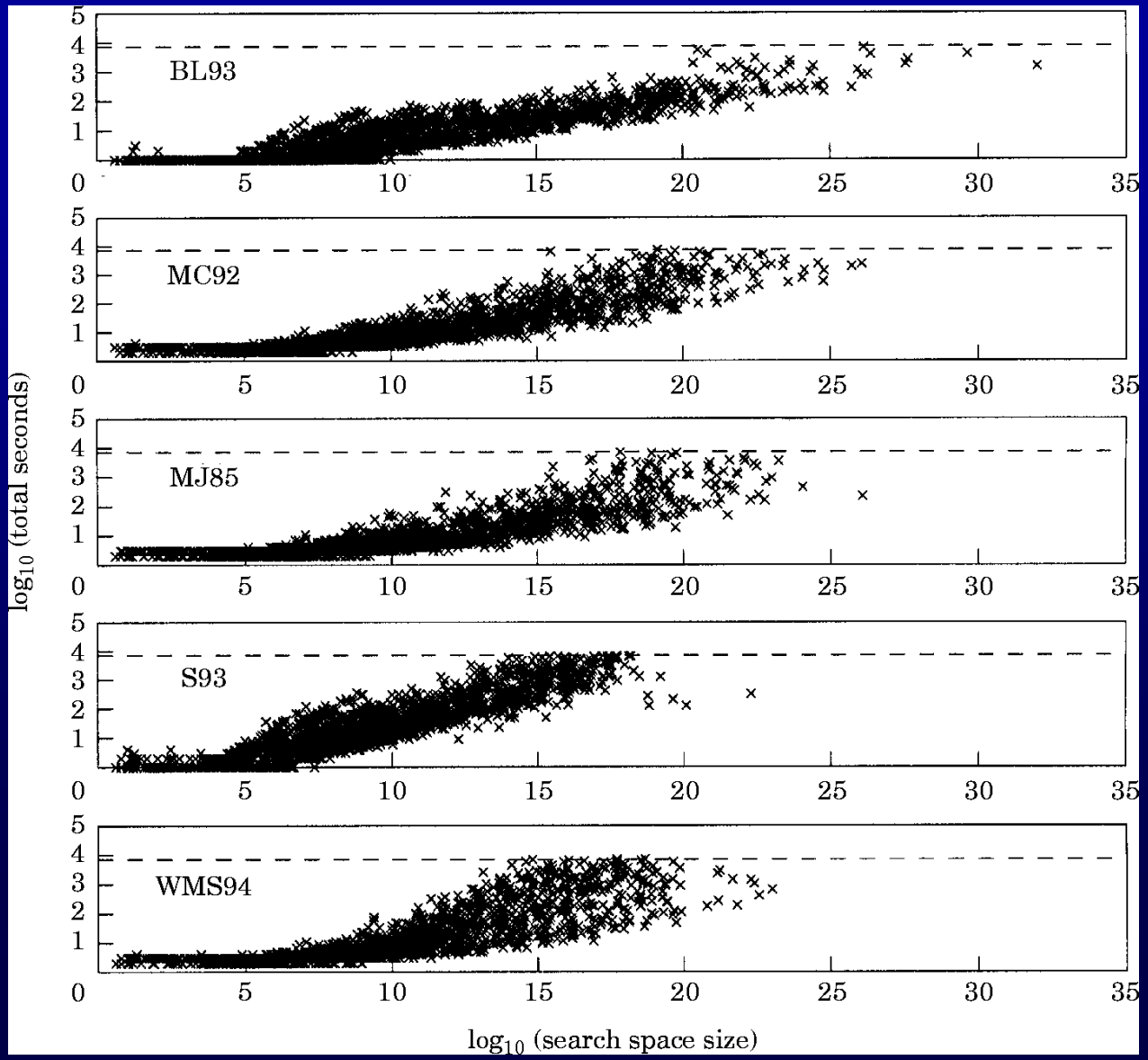
We have explored several alternative forms of the lower bound (Lathrop & Smith, 1994). Our current version, denoted $\mathrm{lb}(\mathscr{T})$, is:

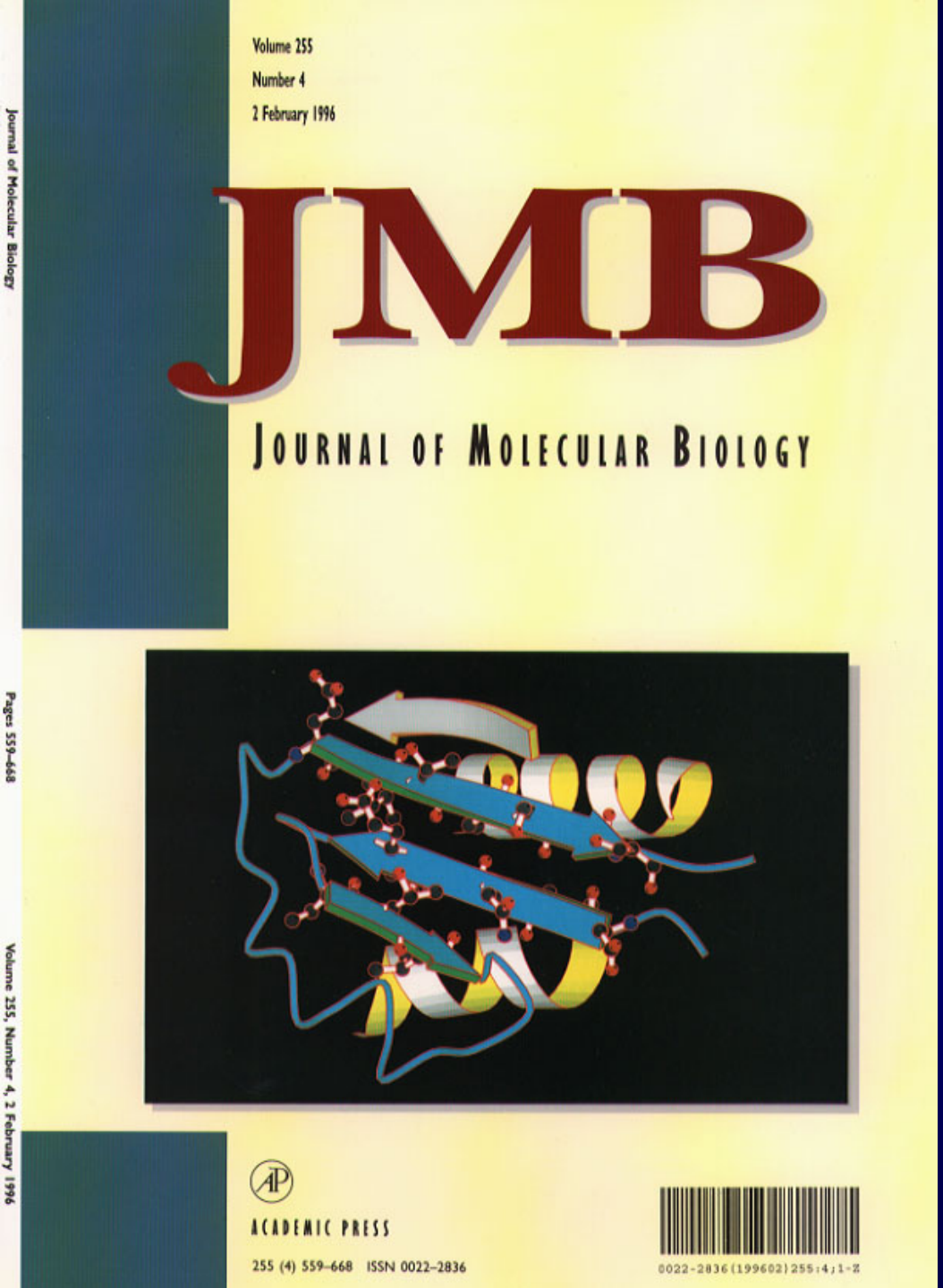
$$\min_{\mathbf{t}\in\mathscr{T}} f(\mathbf{t}) \geqslant \mathrm{lb}(\mathscr{T})$$

$$= \min_{\mathbf{t}\in\mathscr{T}} \sum_{i} \left[g_1(i, t_i) + g_2(i-1, i, t_{i-1}, t_i) \right.$$

$$\left. + \min_{\substack{\mathbf{u}\in\mathscr{T} \\ l_j^{\max}=+\infty}} \sum_{|j-i|>1} \tfrac{1}{2} g_2(i, j, t_i, u_j) \right] \qquad (5)$$

The time required to find the global minimum is shown on log-log axes as a function of search space size.





"Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions"

Lathrop and Smith

J. Mol. Biol. 255(1996)641-665