TABLE 2 HPLC analysis of the oligopeptides produced by translation

	tRNA	Charge	Full-length products (16-mer) ^{35}S (d.p.m.)	%	Termination products (8-mer) ^{35}S (d.p.m.)	%	Frame shift products (polar) ^{35}S (d.p.m.)	%
			mRNA containing the UAG codon					
1	none	—	799	4	20,548	96	76	—
2	CUA	none	821	4	19,291	96	29	—
3	CU(*iso*-dG)	none	524	3	17,844	97	73	—
4	CUA	iodo-Tyr	15,747	67	7,617	33	90	—
5	CU(*iso*-dG)	iodo-Tyr	1,725	9	18,153	91	47	—
			mRNA containing the (*iso*-C)AG codon					
6	none	—	1,145	3	9,195	25	25,832	71
7	CUA	none	967	4	4,597	17	21,749	80
8	CU(*iso*-dG)	none	1,023	4	3,400	14	19,078	81
9	CUA	iodo-Tyr	853	3	6,722	24	20,984	73
10	CU(*iso*-dG)	iodo-Tyr	17,754	91	1,464	8	246	1

Aliquots obtained directly from incubation mixtures were diluted with carrier peptides and injected onto a Vydac C4 column equilibrated in 0.1% trifluoroacetic acid in H_2O/CH_3CN (3:1). Both hydrophobic and hydrophilic peptide, the latter arising from frameshifting, are quantified in this way. Gradient elution with 0.1% trifluoroacetic acid in CH_3CN (25–55% CH_3CN in 60 min) resolved the 8- and 16-mer peaks, eluting at 30–45% CH_3CN. The low read-through of the UAG stop codon seen with non-standard tRNA is expected from the minor tautomeric form of *iso*-G (ref. 27).

translation products, release factors that bind UAG in competition with the semi-synthetic tRNA must be removed or inactivated. However, this would require that UAG triplet stop signals be removed to avoid continued translation after frameshifting. Although it may be possible to fulfil both of these requirements *in vitro*, they are essentially unattainable *in vivo*. Thus for the rearrangement strategy, the low yields of protein obtained from cell-free translation systems cannot be overcome simply by moving the system into a living cell.

Our demonstration that a ribosome can efficiently translate a 65th codon is the first of three breakthroughs required for incorporation of non-standard amino acids into proteins biosynthesized *in vivo*. The second is that plasmids containing a third base pair must be copied and transcribed *in vivo* with reasonable fidelity; the base pair between diaminopyrimidine and xanthosine[2] may be best suited in this regard. Given faithful replication and transcription, a total of 216 codons will be accessible as a result of the introduction of a third base pair. The third and more demanding requirement is for non-standard aminoacyl tRNA synthetases to be engineered which specifically couple a non-standard amino acid to a non-standard tRNA[21,22] □

Received 7 November 1991; accepted 13 February 1992.

1. Switzer, C. Y., Moroney, S. E. & Benner, S. A. *J. Am. chem. Soc.* **111,** 8322–8323 (1989).
2. Piccirilli, J. A., Krauch, T., Moroney, S. E. & Benner, S. A. *Nature* **343,** 33–37 (1990).
3. Bain, J. D., Glabe, C. G., Dix, T. A., Chamberlin, A. R. & Diala, E. S. *J. Am. chem. Soc.* **111,** 8013–8014 (1989).
4. Bain, J. D., Wacker, D. A., Kuo, E. E. & Chamberlin, A. R. *Tetrahedron* **14/15,** 2389–2400 (1991).
5. Hecht, S. M., Alford, B. L., Kuroda, Y. & Kitano, S. *J. biol. Chem.* **253,** 4517–4520 (1978).
6. Roesser, J. R., Chorghade, M. S. & Hecht, S. M. *Biochemistry* **25,** 6361–6365 (1986).
7. Baldini, G., Martoglio, B., Schachenmann, A., Zugliani, C. & Brunner, J. *Biochemistry* **27,** 7951–7959 (1988).
8. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. & Schultz, P. G. *Science* **244,** 182–188 (1989).
9. Robertson, S. A., Ellman, J. A. & Schultz, P. G. *J. Am. chem. Soc.* **113,** 2722–2729 (1991).
10. Kozak, M. *Nature* **308,** 241–246 (1984).
11. Mantsch, H. H. *et al. Biochemistry* **14,** 5593–5601 (1975).
12. McBride, L. J., Kierzek, R., Beaucage, S. L. & Caruthers, M. H. *J. Am. chem. Soc.* **108,** 2040–2048 (1986).
13. Bain, J. D. *et al. Biochemistry* **30,** 5411–5421 (1991).
14. Fluck, M. M., Salser, W. & Epstein, R. H. *Molec gen. Genet.* **151,** 137–149 (1977).
15. Konecki, D. S., Aune, K. C., Tate, W. & Caskey, C. T. *J. biol. Chem.* **252,** 4514–4520 (1977).
16. Smrt, J., Kemper, W., Caskey, T. & Nirenberg, M. *J. biol. Chem.* **245,** 2753–2757 (1970).
17. Valle, R. P. C. & Morch, M.-D. *FEBS Lett.* **235,** 1–15 (1988).
18. Craigen, W. J. & Caskey, C. T. *Cell,* **50,** 1–2 (1987).
19. Tsuchihashi, Z. & Kornberg, A. *Proc. natn. Acad. Sci. U.S.A.* **87,** 2516–2520 (1990).
20. Jukes, T. J., Bessho, Y., Ohama, T. & Osawa, S. *Nature* **352,** 575 (1991).
21. Fersht, A. R. *Biochemistry* **26,** 8031–8037 (1987).
22. Kast, P. thesis ETH, Zurich (1991).
23. Roberts, J. W. & Carbon, J. *J. biol. Chem.* **250,** 5530–5541 (1975).
24. Lyttle, M. H., Wright, P. B., Sinha, N. D., Bain, J. D. & Chamberlin, A. R. *J. org. Chem.* **56,** 4608–4615 (1991).
25. Bain, J. D., Wacker, D. A., Kuo, E. E., Lyttle, M. H. & Chamberlin, A. R. *J. org. Chem.* **56,** 4615–4625 (1991).
26. Fersht, A. R. & Dingwall, C. *Biochemistry* **18,** 2627–2631 (1979).
27. Sepiol, J., Kazimierczuk, Z. & Shugar, D. *Z Naturf.* **31c,** 361–370 (1976).

ACKNOWLEDGEMENTS. We thank S. Hecht and J. Knowles for comments on the manuscript. This work was supported by the Swiss Nationalfond, Sandoz, the ETH and the NIH. C.S. was supported by a fellowship from the NSF.

The dead-end elimination theorem and its use in protein side-chain positioning

Johan Desmet*, Marc De Maeyer†, Bart Hazes‡ & Ignace Lasters†

* Interdisciplinary Research Center, KU Leuven Campus Kortrijk, B8500 Kortrijk, Belgium
† Corvas International NV, Jozef Plateaustraat 22, B9000 Gent, Belgium
‡ BIOSON Research Institute, Department of Chemistry, Chemical Physics, RU Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands

THE prediction of a protein's tertiary structure is still a considerable problem because the huge amount of possible conformational space[1] makes it computationally difficult. With regard to side-chain modelling, a solution has been attempted by the grouping of side-chain conformations into representative sets of rotamers[2–5]. Nonetheless, an exhaustive combinatorial search is still limited to carefully identified packing units[5,6] containing a limited number of residues. For larger systems other strategies had to be developped, such as the Monte Carlo Procedure[6,7] and the genetic algorithm and clustering approach[8]. Here we present a theorem, referred to as the 'dead-end elimination' theorem, which imposes a suitable condition to identify rotamers that cannot be members of the global minimum energy conformation. Application of this theorem effectively controls the computational explosion of the rotamer combinatorial problem, thereby allowing the determination of the global minimum energy conformation of a large collection of side chains.

The potential energy of a protein system consisting of a set of residue side chains i each in a particular rotameric state r and embedded in a template of fixed atoms (for example, main-chain atoms) can be written as

$$E_{\text{global}} = E_{\text{template}} + \sum_i E(i_r) + \sum_i \sum_j E(i_r j_s); \quad i<j \qquad (1)$$

where E_{template} is the template self energy, $E(i_r)$ the potential energy of the side chain atoms of the rotamer i_r in the force field of the template, and $E(i_r j_s)$ the nonbonded pairwise interaction energy between rotamers i_r and j_s. The calculation of the terms $E(i_r)$ and $E(i_r j_s)$ is computationally quite feasible as their numbers grow linearly and quadratically, respectively, as a function of the number of residues. On the basis of these energies, a condition can be formalized, allowing identification

of rotamers i_r that are absolutely incompatible with the global minimum energy conformation (GMEC). In a combinatorial search to identify the GMEC, conformations comprising such rotamers can be qualified as dead ending. Therefore, this condition will be referred to as the dead-end elimination (DEE) theorem: if for a couple of rotamers (i_r, i_t) the following inequality holds true

$$E(i_r)+\sum_j \min_s E(i_r j_s) > E(i_t)+\sum_j \max_s E(i_t j_s); \quad i \neq j \qquad (2)$$

then i_r is incompatible with the GMEC. In equation (2) the functions $\min_s$ and $\max_s$ are evaluated by searching the lowest and the highest value, respectively, for their argument by cycling over s. This search procedure grows quadratically with the total number of rotamers considered.

The mathematical proof of this theorem readily follows from the definition of the GMEC, which implies that any conformation must necessarily have a higher or equal (in case of degeneracy) energy than the GMEC itself

$$E_{\text{global}} \geq E_{\text{GMEC}} \qquad (3)$$

Expressing the right-hand side of this equation making use of

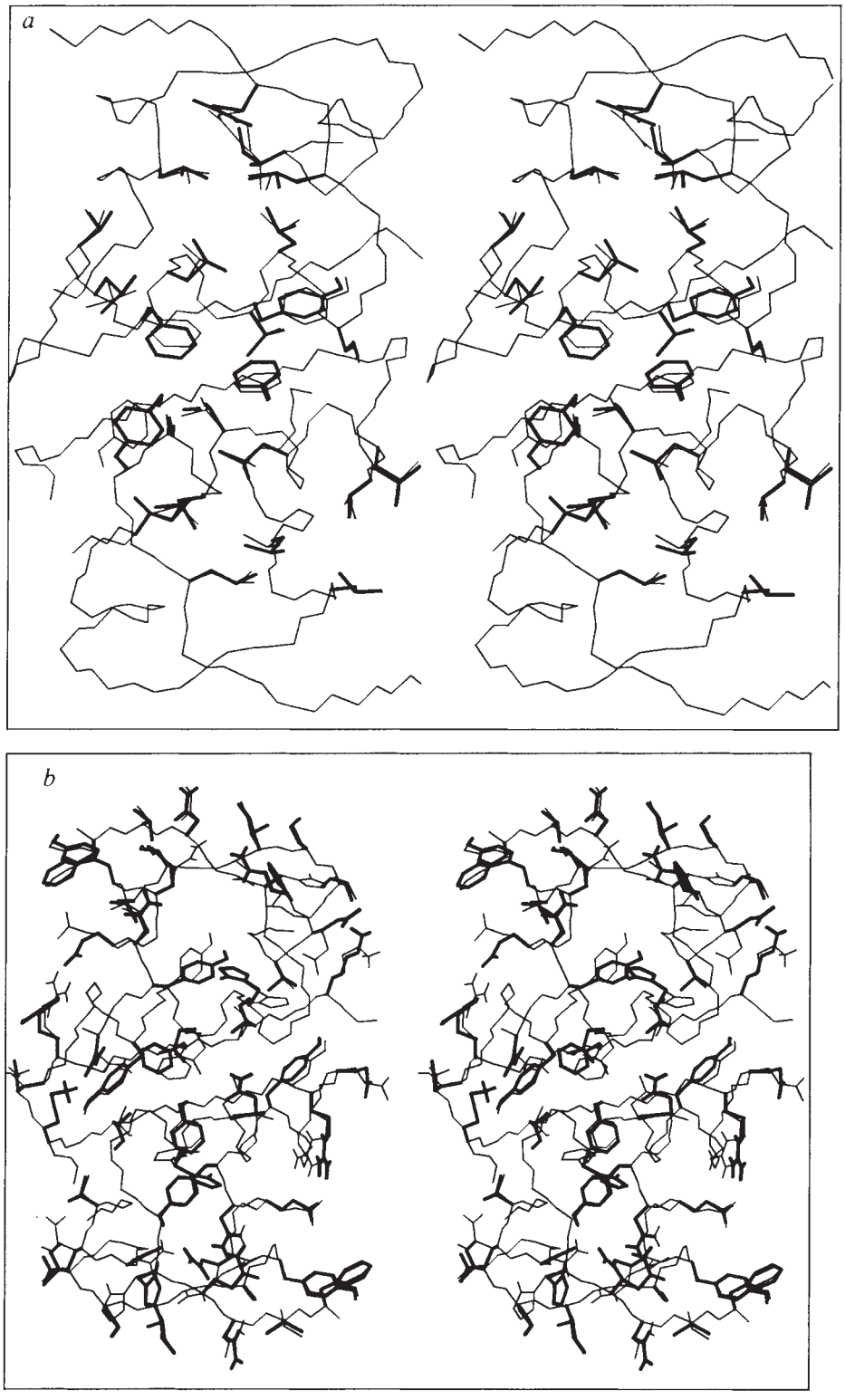

FIG. 1 Stereoscopic view of the insulin dimer (coordinates are obtained from the Brookhaven protein databank[11], code 3ins). The other two dimers in the insulin hexamer were included in the prediction procedure to allow a correct modelling of subunit–subunit interfaces. One dimer contains 102 amino acids, 26 of which are Gly, Ala, Pro or Cys involved in a disulphide bridge. These residues do not contain freely rotatable heavy atoms, and were therefore considered as being part of the template. The other 76 residues were simultaneously repositioned. An extended version of the rotamer library[5] was used. For all aromatic residues, rotamers created by taking steps of size equal to the standard deviation[5] above and below dihedral angles χ_1 and χ_2 were added to the library. In this way the library, which is considerably larger than those used in previous studies[5,7,8], contained 363 rotamers to model 16 residue types. This library is available on request. The number of rotamers for each of the rotatable side chains are Val and Asp, 3; Leu, 4; Ile, 5; Asn, 6; Glu, 7; Ser and Thr, 9; Gln, 10; Met, 21; Phe and Tyr, 36; Lys, 51; His and Trp, 54; Arg, 55. Side-chain placement was done in the presence of all H atoms using a molecular mechanics model[12]. The backbone of the complete dimer as well as the residues exposing less (*a*) or more (*b*) than 5% of their maximum possible accessible surface area are shown. The predicted conformations are represented in bold lines. To prevent any bias masking the outcome of our method, the modelled structure was not energy-minimized. The r.m.s. deviation between the observed and the predicted heavy side-chain atoms (all residues excluding Gly, Ala, Pro and Cys residues) is 1.69 Å (76 residues). The r.m.s. value calculated with the same critieria for the less than 5% and more than 5% exposed residues are 0.61 Å (24 residues) and 1.96 Å (52 residues), respectively.

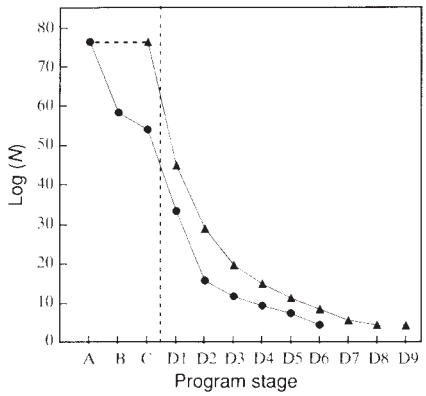

FIG. 2 Evolution of the logarithm of total number of possible rotamer combinations (N) as a function of the program stage in positioning the side chains of the insulin dimer. Program states are as follows: A, start of the program; B, after elimination of rotamers that are clashing with the surrounding template using a threshold energy value of 30 kcal mol^{-1}; C, after elimination of rotamers that are clashing with all rotamers of some residue using the same threshold; D1–D9, interations of the DEE procedure. ▲, Evolution of N as a function of program stage, excluding stages B and C. The total required computing time was 5.8 h on an IRIS 4D/310 computer of which 5.3 h was spent in the computation of the pairwise rotamer interaction energies (780,000 pairs). The DEE calculations took only 14 min. ●, Evolution of N as a function of program stage including stages A and B. The total computing time required was 1.8 h of which 1.2 h was consumed in rotamer interaction computation and only about 3 min in DEE calculations.

equation (1), isolating the terms for a particular residue of interest i and denoting GMEC-compatible rotamers with the subscript g, we obtain

$$E_{\mathrm{GMEC}} = E_{\mathrm{template}} + E(i_g) + \sum_j E(i_g j_g) + \sum_j E(j_g) + \sum_j \sum_k E(j_g k_g); \quad j, k \neq i; \quad j < k \tag{4}$$

If the residue of interest, i, adopts an undefined rotamer conformation i_t, eventually different from i_g, we can write an analogous expression for the energy of the resulting conformation

$$E_{\mathrm{global}} = E_{\mathrm{template}} + E(i_t) + \sum_j E(i_t j_g) + \sum_j E(j_g) + \sum_j \sum_k E(j_g k_g); \quad j, k \neq i; \quad j < k \tag{5}$$

Substitution of equations (4) and (5) into equation (3) gives

$$E(i_t) + \sum_j E(i_t j_g) \geq E(i_g) + \sum_j E(i_g j_g); \quad i \neq j \tag{6}$$

Since for all i, j $(i \neq j)$,

$$\max_s E(i_t j_s) \geq E(i_t j_g) \tag{7}$$

and

$$\min_s E(i_g j_s) \leq E(i_g j_g) \tag{8}$$

it follows that

$$E(i_t) + \sum_j \max_s E(i_t j_s) \geq E(i_g) + \sum_j \min_s E(i_g j_s); \quad i \neq j \tag{9}$$

Now, each rotamer i_r that satisfies the dead-end criterion of equation (2) must, in view of equation (9), necessarily obey the inequality

$$E(i_r) + \sum_j \min_s E(i_r j_s) > E(i_g) + \sum_j \min_s E(i_g j_s); \quad i \neq j \tag{10}$$

This implies that $i_r \neq i_g$ thereby proving the DEE theorem.

By the same reasoning, the DEE theorem can be extended to rotamer pairs or, in general, to multi-residue rotamer combinations. Hereafter, only the extension towards pairs is given. The latter is analogous to the single rotamer criterion, and so is the extension towards multiple rotamer entities. Defining a specific energy term for the rotamer pair $[i_r j_s]$, $\varepsilon([i_r j_s])$ as

$$\varepsilon([i_r j_s]) = E(i_r) + E(j_s) + E(i_r j_s); \quad i \neq j \tag{11}$$

and the interaction energy between this pair and another rotamer k_t, $\varepsilon([i_r j_s] k_t)$ as

$$\varepsilon([i_r j_s] k_t) = E(i_r k_t) + E(j_s k_t); \quad i, j \neq k \tag{12}$$

then, a rotamer pair $[i_r j_s]$ is dead ending if there exists a $[i_u j_v]$ pair for which

$$\varepsilon([i_r j_s]) + \sum_k \min_t \varepsilon([i_r j_s] k_t) > \varepsilon([i_u j_v]) + \sum_k \max_t \varepsilon([i_u j_v] k_t); \quad i, j \neq k \tag{13}$$

The proof for this extended DEE theorem can be derived by analogy to the proof of the DEE theorem for single rotamers. Application of the DEE theorem to rotamer pairs grows cubically with the number of rotamers and is therefore the slowest step of the dead-end elimination procedure, once the terms $E(i_r)$ and $E(i_r j_s)$ have been calculated.

The DEE theorem was implemented in a combinatorial substitution algorithm developed in the Brugel[9] package. We report here an in depth study for the insulin dimer using an extended version of the Ponder and Richards rotamer library[5] (Fig. 1). The initial number of rotamer combinations (N) for the 76 residues to be positioned in the insulin dimer equals 2.7×10^{76} (Fig. 2, stage A). The power of the DEE theorem is exemplified in the major reduction of this huge rotamer conformational space, especially when applied iteratively (Fig. 2, stages D1 to

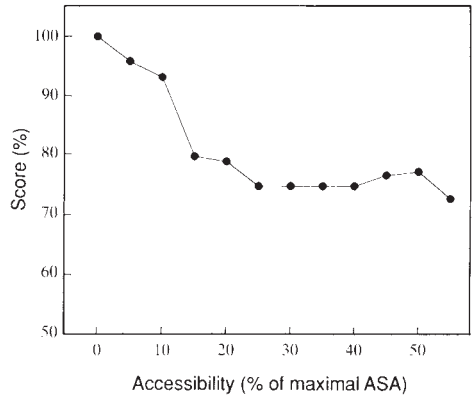

FIG. 3 Score for correct prediction of residue side-chain orientation for residues having an accessibility below the indicated value (expressed as per cent). Residues are considered to be correctly predicted if both their χ_1 and χ_2 angles fall in the same dihedral class as observed for the X-ray structure. The accessibility of a residue side chain is defined as the ratio of the ASA in the protein versus the maximal ASA in an extended conformation in a tripeptide and expressed in per cent.

D9). In each iteration stage, two operations are carried out: (1) the DEE theorem is applied to single rotamers until no more dead ending rotamer can be identified; (2) the generalized DEE theorem (equation (13)) is then applied to all possible rotamer pairs and dead ending pairs are flagged. In the next iterative stage, these flagged dead ending pairs can lead to the further elimination of rotamers because: (1) pairs that have been flagged are discarded from the $\sum_j \min_s E(i,j_s)$ computation (equation (2)), which possibly augments the left-hand side of the DEE criterion; and (2) if for a given rotamer all its rotamer pairs with some other residue have been flagged, this rotamer evidently will be dead ending too. The iteration ends when no dead ending rotamers or pairs can be identified. For the insulin dimer this resulted in nine stages (Fig. 2, D1 to D9, upper curve) whereby N was reduced from 2.7×10^{76} to only 7,200, a reduction of N by 72 orders of magnitude.

The number of computations can be considerably reduced by first eliminating rotamers that are clearly incompatible with the given template or with all rotamers of some other residue. Using a 30 kcal mol^{-1} threshold value, 44% and 6% of the rotamers were eliminated because of incompatibility with template and surrounding side chains, respectively. This elimination reduced the initial N from 2.7×10^{76} to 6.2×10^{53} (Fig. 2, stages B and C). Applying the DEE theorem further reduced N to 10,800 in six iteration cycles (Fig. 2).

After applying the DEE theorem in both experiments, only seven residues had more than one rotamer left. Their side chains were modelled by the following 'add on' mechanism. In each addition step, the rotamers of the added residue were combined with the group of allowed rotameric combinations from the previous step, eventually leading to the rejection of combinations by applying the generalized DEE theorem for multi-residue systems. At each step no more than 23 rotameric combinations had to be stored in memory. The lowest energy conformation obtained in the final addition step was retained as the ultimate structure and compared with the X-ray-determined structure as is illustrated in Fig. 1. Both experiments yielded the same final structure. This must be the absolute GMEC for the given rotamer library and energy function, because in the first experiment no user-defined threshold value was used and also the DEE theorem can only identify GMEC-incompatible rotamers.

Residues are considered to be correctly predicted if both their χ_1 and χ_2 angles fall in the same dihedral class as observed in the X-ray structure. According to this criterion, 55 out of the 76 residues studied were correctly predicted, with average deviations to the X-ray model of $\Delta\chi_1 = 8.9°\pm7.0°$ and $\Delta\chi_2 = 12.8°\pm9.2°$. A clear correlation was observed between the correctness of the prediction and the extent of burying (Figs 1 and 3). The residues with a solvent accessible surface area (ASA) up to 10% of their maximum ASA (29 cases) are predicted with an accuracy of 93%. This sharply drops (Fig. 3) to a rather constant level of about 72% when including residues with higher accessibilities. The occurrence of this drop at the level of 10% nicely corresponds to the 5–8% value proposed by Miller *et al.*[10] in their definition of buried residues.

Finally, to assess the prediction power of the DEE theorem for larger systems, a blind test was done on the third domain of Limulus polyphemus subunit type II haemocyanin (resolution = 2.2 Å, R factor = 17.4%). Only backbone coordinates of this structure were provided by one of the authors (B.H., manuscript in preparation). We predicted 194 amino-acid side chains and then compared them with the X-ray model. An overall prediction score of 71% was obtained, with average deviations on χ_1 and χ_2 angles for the correctly predicted side chains of $\Delta\chi_1 = 7.6°\pm6.5°$ and $\Delta\chi_2 = 15.0°\pm14.1°$, respectively. Visual inspection suggested that many of the errors were caused by incorrectly placed salt bridges and hydrogen-bonded interactions involving exposed polar groups. The lower score for these residues can be reasonably attributed to reduced packing constraints and to the absence of water molecules. Modification of the force field is therefore expected to improve the predictions.

We have proposed a theorem that allows, for a given library of rotamers and energy function, an effective search for the GMEC of a large collection of side chains embedded in a given fixed template. In addition to its theoretical interest, we believe that the DEE theorem can be valuable in homology modelling and in predicting side-chain rearrangements, provided a reasonably accurate template is available. Moreover, the high performance of the algorithm now brings into reach intriguing applications, such as flexible substrate-protein docking (whereby protein side chains are allowed to adapt to the substrate molecule being docked) and protein modelling involving main-chain adjustments. □

Received 25 November 1991; accepted 3 February 1992.

1. Levinthal, C. *J. chim. Phys.* **65,** 44–45 (1968).
2. McGregor, M. J., Islam, S. A. & Sternberg, M. J. E. *J. molec. Biol.* **198,** 295–310 (1987).
3. Janin, J., Wodak, S., Levitt, M. & Maigret, B. *J. Molec. Biol.* **125,** 357–386 (1978).
4. James, M. N. G. & Sielecki, A. R. *J. molec. Biol.* **163,** 299–361 (1983).
5. Ponder, J. W. & Richards, F. M. *J. molec. Biol.* **193,** 775–791 (1987).
6. Lee, C. & Levitt, M. *Nature* **352,** 448–451 (1991).
7. Holm, L. & Sander, C. *J. molec. Biol.* **218,** 183–194 (1991).
8. Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. *J. biomolec. struct. Dyn.* **8,** 1267–1289 (1991).
9. Delhaise, P., Bardiaux, M. & Wodak, S. *J. J. molec. Graph.* **2,** 103–106 (1984).
10. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. *J. molec. Biol.* **196,** 641–656 (1987).
11. Berenstein, F. C. *et al. J. molec. Biol.* **112,** 535–542 (1977).
12. Brooks, B. R. & Karplus, M. *J. comput. Chem.* **4,** 187–217 (1983).

ACKNOWLEDGEMENTS. We thank W. Hol for discussion. J.D. thanks F. Van Cauwelaert for scientific advice, S. Wodak for introduction to the modelling field and the Research Foundation of Leuven and the FGWO for financial support.