# Different molecular size scaling regimes for inner and outer regions of proteins

Gustavo A. Arteca
*Département de Chimie et Biochimie, Laurentian University, Ramsey Lake Road, Sudbury, Ontario, Canada P3E2C6*

(Received 20 December 1995)

We study the rough statistical features of size scaling in protein backbones by using molecular descriptors associated with their central and external regions. By using a diverse set of experimental structures, we show that the mean radius of gyration and the span of backbones scale differently, leading to a ratio of ''inner'' and ''outer'' scaling size exponents of $\nu_i/\nu_o \approx 0.8$. The span of average proteins is found to scale with the number of amino acids as $R \sim n^{\nu_o}$, with $1/2 < \nu_o < 3/5$, thus providing a measure of the ''locally'' swollen character of the backbone near the exterior. The result holds for all classes of proteins, including those with the most compact cores. [S1063-651X(96)12009-2]

PACS number(s): 87.15.He, 82.20.Wt, 05.90.+m

The selective packing of amino acid residues according to hydrophobicity is a key factor determining the three-dimensional structure of proteins. Globular proteins exhibit a slight dominance of hydrophobic amino acid residues (55%), whereas soluble proteins have an even population of polar and hydrophobic residues [1]. These two groups of proteins make up the large majority of those for which x-ray structures are available. As a result, the ''average'' configuration of the known protein native states is expected to include a hydrophobic core and an exterior dominated by hydrophilic residues [2]. Whereas each protein native state is a singular conformation evolved to serve a specific biological function, it is important to test whether the distribution of native states over a large database exhibits some defined rough ''universal'' statistical features. In this work we tackle this issue: the characterization of average properties common to a large number of protein configurations.

The relation between the total number of residues ($n$) and the change in molecular size when moving from the interior to the exterior of the globule is an important piece of information about the protein's configurational state. Understanding the interrelation between size, compactness, and hydrophobicity is central towards unraveling the mechanism and reaction intermediates of the folding pathway [3–6, and references therein]. Recent work has commented on the possible existence of power-law scaling in a subclass of compact proteins. There is evidence that the mean size of the smallest globular proteins resembles that of collapsed polymers [7,8]. These results suggest that it is indeed possible to apply concepts from scaling theory of polymers [9] to the study of the medium-size biopolymers. However, it must be noted that the ''collapsed state'' is not the standard configurational state of most proteins. Results indicate a dependence of protein compactness on chain length [8]. Yet, nothing is known on how the scaling regimes of molecular sizes for inner and outer regions compare. Here, we assess their relation by studying the change in molecular size scaling across the protein (from inner to outer regions). We focus on two distinct geometrical descriptors adapted to study different sections of experimental protein backbones. Contrary to the behavior in random homo- and heteropolymers, we show that the shape descriptors of protein native states present different scaling behavior depending on whether they are associated with a ''mean size'' or to an ''external size.'' The result provides a quantitative comparison of the swollen vs compact character of sections of a protein backbone.

Backbones are specified by the positions of $n$ $\alpha$-carbons (one per residue), $\{\mathbf{r}_i, i=1,2,\ldots,n\}$, as deposited in the Brookhaven Protein Data Bank (PDB) [10]. [The centroid of the $\alpha$-carbon backbone is taken as the origin.] For simplicity, we characterize the backbone size with only two geometrical descriptors: (i) the radius $R$ of the smallest sphere (centered at the centroid) which encloses completely the backbone (the ''span''), and (ii) the instantaneous radius of gyration $R_G$

$$R = \max_{\{i\}} r_i, \quad r_i = \|\mathbf{r}_i\|; \quad R_G^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2, \quad R_G \leq R. \tag{1}$$

[Another definition of the span, using an enclosing box rather than a sphere, can also be used [11]. This approach leads to the same conclusions.] The span $R$ is determined by a single residue (the farthest from the centroid), and it represents the state of the outer layers. In contrast, the radius of gyration takes into account all residues and gives a mean size. Since normally the value of $R_G$ will be controlled by the inner layers, we use $R_G$ as a measure of ''internal size.'' A simple power-law scaling with the number of residues is assumed, $R_G \sim n^{\nu_i}$ and $R \sim n^{\nu_o}$ with two distinct exponents for the internal ($\nu_i$) and the external ($\nu_o$) radii. We test whether a relation

$$R_G \approx aR^g, \quad g = \nu_i/\nu_o, \tag{2}$$

is found in proteins, and whether the random-polymer result ($g \approx 1$) is valid for their native states.

From the behavior of *random linear polymers*, the following results are known: (i) All molecular size functions scale equally in terms of polymer length (or $n$ in our case), i.e., $g=1$ [9]. (ii) If the polymer is in a $\theta$ solvent (i.e., an ''ideal'' poor solvent where repulsive and attractive monomer-monomer interactions balance each other), we expect $\nu_i = \nu_o = 1/2$ [9]. (iii) If the polymer is in a ''good'' solvent, the chains adopt swollen conformations and the size exponents are larger $\nu_i = \nu_o \approx 3/5$. (The actual value is $0.588 \pm 0.002$ [12]; the exponent 3/5 corresponds to a mean-field approach.) (iv) If the polymer is in a very poor solvent, it is expected to appear in collapsed, maximally compact, conformations [13], where $\nu_i = \nu_o = 1/3$.

In the case of general proteins, it is difficult to establish the nature of the dominant conformations from standard correlations between the number of residues $n$ and the radius of

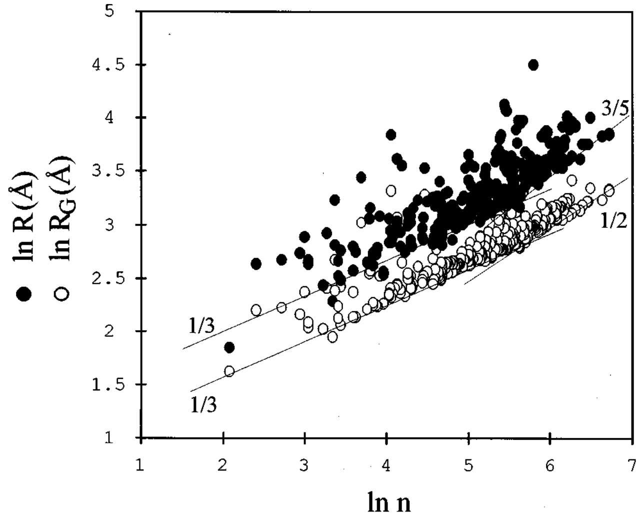

FIG. 1. Distribution of backbone spans ($R$) and radii of gyration ($R_G$) for the working set of 373 proteins as a function of the number of amino acids ($n$). [The numbers 1/3, 1/2, and 3/5 indicate the limiting slopes associated with the $\nu_i$ and $\nu_o$ exponents expected in collapsed, ideal, and swollen polymers, respectively.]

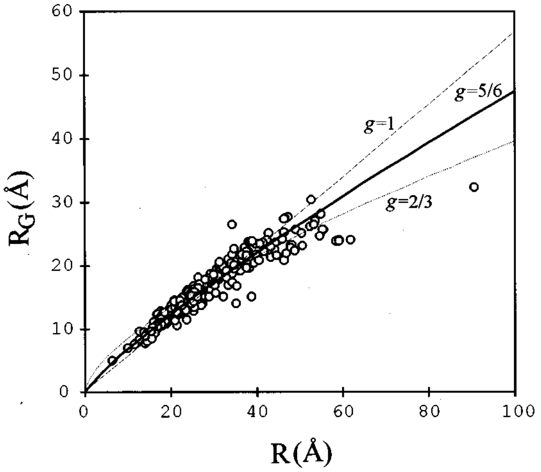

FIG. 2. Correlation between size descriptors for inner ($R_G$) and outer ($R$) layers in the set of 373 proteins. [The correlation is nonlinear for proteins with $R > 30$ Å ($n > 300$). The curves indicated fit various scaling exponents $g = \nu_i / \nu_o$.]

gyration. Using a small set of proteins, Dewey suggested the universal validity of an exponent $\nu_i \approx 1/3$ [7]. Recently, we have shown with a much larger set of structures (selected to avoid biases in secondary structure) that the dispersion in $R_G$ vs $n$ is too large to derive a reliable scaling exponent for all proteins [8,14]. However, in the subclass of proteins with the *smallest backbone radius of gyration* within a fixed range of monomer numbers $n$, a well-defined scaling behavior is found with an exponent $\nu_i \approx 0.4$ [8]. (The conclusions are the same if the actual volumes of individual residues are taken into account in an alternative definition of ‘‘compactness’’ [15,16].) Only in the case of the *short* maximally compact proteins ($n \leqslant 300$), an exponent close to the collapsed polymer regime is found ($\nu_i \approx 0.34 \pm 0.05$) [8]. In summary, whereas there is evidence that the *mean size* (as measured by $R_G$) of the smallest globular proteins resembles that of collapsed polymers, the situation for the average protein native state is not so simple. All recent studies have centered on properties of the radius of gyration. Below, we show that important information can be extracted by analyzing the behavior of $R$ for average native states, as well as for special classes of proteins (e.g., those with the global constraint of being maximally compact).

We have analyzed $R_G$ and $R$ in a working set of 373 proteins. The set includes proteins of various lengths, composition, and structural content [8]. It has been chosen to maximize diversity. Nearly identical proteins have been excluded. The set includes no structural bias and we believe it should properly convey the rough statistical features of the average known native states. (See Ref. [8] for the mean molecular shape properties of the proteins in the set.)

Figure 1 shows the dependence of $R_G$ and $R$ on the monomer number $n$. It is clear that a simple $n$ scaling cannot be assigned to the proteins, except for those that minimize the radii. For a qualitative reference, we indicate the limiting lines associated with the random-polymer scaling exponents that appear to fit best the proteins with smallest $R_G$ and $R$ values. These exponents are illustrative and they may not apply, in principle, to the entire protein set. However, Fig. 1 suggests that $R_G$ and $R$ may not scale in the same manner as a function of $n$, for large $n$ values.

Whereas Fig. 1 shows a large dispersion in terms of $n$, Fig. 2 shows that the interrelation between radii is better defined. For the set of 373 proteins, a log-log correlation gives

$$R_G \approx (1.2 \pm 0.1) R^{(0.80 \pm 0.02)} \quad \text{(radii in angstroms)}, \quad (3)$$

with 95% confidence errors (and a correlation coefficient of $\mathcal{C} = 0.961$). [Pseudolinear regression models $R_G$ vs $R^g$, where $g$ maximizes $\mathcal{C}$, lead to comparable results: $R_G \approx (1.6 \pm 0.1) R^{(0.71 \pm 0.02)}$, with $\mathcal{C} = 0.947$.] The linear model [$g = 1$ in Eq. (2)] can be readily discarded because of its poor correlation $R_G \approx 0.57\, R$, with $\mathcal{C} = 0.858$ (dashed line in Fig. 2). The linear correlation does, however, give us a bound to the radius of gyration of a protein in terms of its span: $R_G \leqslant 0.63 R$. [The significance of correlation (3) has been tested by also evaluating the exponent $g$ in a ‘‘control set’’ of linear polymers with comparable numbers of monomers. We have generated a series of random chain conformations with two characteristics [14]: (i) a constant step, similar to the distance between $\alpha$-carbons in proteins ($l = 3.8$ Å), and (ii) variable excluded volume interaction. In the limit of no-excluded volume, the correlation between $R_G$ and $R$ produces an exponent $g = 0.96 \pm 0.02$, with $\mathcal{C} = 0.999$ (and $\nu_i = 0.500 \pm 0.001$). For large excluded volume, we find $g = 0.95 \pm 0.05$, with $\mathcal{C} = 0.998$ (and $\nu_i = 0.57 \pm 0.01$). These results agree with the expected limit $g \approx 1$ for random polymers and suggest that correlation (3) is a meaningful deviation associated with the occurrence of special structural features in proteins. In addition, note that the correct statistical behavior for homopolymers (i.e., $g \approx 1$) was achieved with a control set of off-lattice model chains with only $n \leqslant 500$ monomers. This suggests that the chain lengths for proteins within our set ($n < 824$) should be long enough to extract a qualitative scaling behavior.]

In order to provide an interpretation to the correlation in Eq. (3), a number of fixed scaling models have also been explored over the entire set of proteins, in addition to the linear one. From the scaling regimes discussed before for random polymers, only three $g<1$ power laws could be possible: (a) a compact center ($\nu_i=1/3$) and a less compact exterior ($\nu_o=1/2$), i.e., $g=2/3\approx0.67$; (b) a center at intermediate compactness ($\nu_i=1/2$) and a swollen exterior ($\nu_o=3/5$), i.e., $g=5/6\approx0.83$; (c) a compact center ($\nu_i=1/3$) and a swollen exterior ($\nu_o=3/5$), i.e., $g=5/9\approx0.56$. The regression analysis with fixed $g$ exponent leads to comparable results in the above cases

$$R_G\approx(2.68\pm0.03)R^{5/9},\quad \mathcal{C}=0.923, \tag{4a}$$

$$R_G\approx(1.83\pm0.02)R^{2/3},\quad \mathcal{C}=0.946, \tag{4b}$$

$$R_G\approx(1.02\pm0.01)R^{5/6},\quad \mathcal{C}=0.931. \tag{4c}$$

Correlations (4b) and (4c) are superimposed to the experimental results in Fig. 2. The empirical scaling exponent for the entire working set [$g\approx0.80\pm0.02$ in Eq. (3)] appears to be consistent with the model scalings (4b) and (4c). Nevertheless, the agreement is, at best, qualitative. It should *not* be taken as inequivocal indication that the exterior residues resemble chains in a ''good'' solvent. It is also possible that the change in scaling exponent $\nu_i$ reflects a ''surface correction'' to the radius of gyration of a finite polymer. In this case, a behavior such as $n\sim aR_G^3+bR_G^2$ would be expected, but this effect cannot be discriminated with our data.

Note that the functions in Eqs. (4a)–(4c) ''cross'' among themselves and with the simple linear correlation at nearly the same values: $R\approx33\pm2$ Å and $R_G\approx19\pm1$ Å . These radii correspond to proteins with a critical number of residues $n_C\approx300\pm50$. For proteins below the $n_C$ value, the relation between $R$ and $R_G$ is closer to linear, indicating a similar configuration state (residue packing) for sections of the backbone across the globule. Proteins longer than $n_C$ deviate from this behavior and lead inequivocally to $g<1$, suggesting different packing features in the exterior.

The above result indicates that the average configurational state of longer proteins is less compact than the one for shorter proteins. A similar difference in compactness as a function of length had been proposed on different arguments for globular proteins [13]. Our present finding of a change in scaling law at a ''critical'' number of residues $n_C\approx300$ is also consistent with results on the compactness of multidomain proteins. The consensus is that proteins with more than ca. 250 residues usually fold by forming separate domains [17,18] and have nonspheroidal native states [13]. These larger proteins are expected to have a different proportion of hydrophobic residues at the exterior surface with respect to smaller (single-domain) proteins [13]. This distinct behavior could be a factor leading to the difference in inner and outer size scalings observed here. Thus it is possible that our $n_C$ value indicates the beginning of multidomian proteins, or the onset of distinct supersecondary structure only accessible to larger proteins, e.g., $\alpha/\beta$ barrels.

Recently, we have shown that such a change in size-scaling behavior is apparent also in the subclass of proteins with maximal compactness (those with a minimum backbone radius of gyration over a range of residues) [8]. In this latter case, a distinct behavior in $R_G$ vs $n$ is observed for $n<300$ and $n>300$. As Fig. 1 shows, a direct $R_G$ vs $n$ analysis is not feasible for the overall ensemble of proteins. However, a clearer behavior is revealed by the correlation between two distinct geometrical descriptors, such as $R_G$ and $R$. The present results suggest that previous observations on the configurational state of some special proteins (those with compact backbones or minimal volumes) *can be extended to the average state of all native states*. We observe that: (i) in short proteins, the core and the exterior chain section appear to be in similar scaling regimes; (ii) in long proteins, the exterior region appears to be in a regime corresponding to more ''swollen'' chains. Note that the strong deviation from the $g=1$ regime is a clear indication of differential packing across the protein. Nevertheless, it should be noted that there are several mechanisms that can account for the ''swollen'' nature of the farthest residues, with distinct solvation being only one of them. The present results convey a fact but provide no unique interpretation for it.

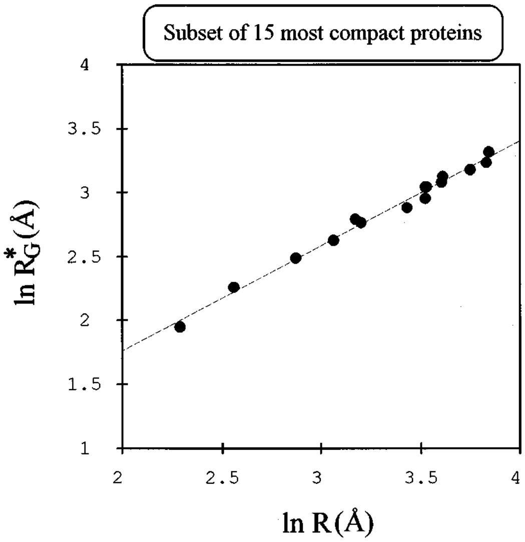

FIG. 3. Nonlinear scaling between size descriptors in the proteins with maximally compact backbones. [The slope indicates the same differential scaling of outer and inner regions found in the average state of all proteins: $R_G^*\sim R^{0.8}$.]

The above results apply to the complete set of 373 proteins. We have also checked the scaling behavior in the subset of proteins with maximal backbone compactness [8]. Here, we select proteins whose backbones have minimal radius of gyration over a ''window'' in the number of residues $n$. That is, we select *one* protein within a given ''bin'' $[n_o,n_o+\Delta n]$, with radius $R_G^*(n_o^*)$

$$R_G^*(n_o^*)=\min_{n\in[n_o,n_o+\Delta n]} R_G(n),\quad n_0\leqslant n_o^*\leqslant n_o+\Delta n, \tag{5}$$

where the $n_o^*$ is the number of residues of the selected protein per bin. We have checked the scaling behavior of size descriptors in the set of the 15 most compact proteins found within the ranges [20,49], [50,99], [100,149], etc., corresponding to $\Delta n=50$. (Proteins that are too short were excluded since their $R_G$ are trivially small. See Ref. [8] regarding the molecular shape properties of the proteins in this ensemble.) The relation between the inner and outer radii for this set of maximally compact proteins is shown in Fig. 3. There is a clear (nonlinear) correlation between the radii. The scaling exponent found is close to the average behavior of the entire working set of 373 proteins

$$R_G^*\approx(1.1\pm0.2)R^{(0.82\pm0.05)},\quad \mathcal{C}=0.995. \tag{6}$$

The virtual coincidence of the $g$ exponents for average proteins and for the most compact ones is a strong indication that native states of *all* folded proteins share the same rough features with respect to packing (in spite of different compositions, secondary structure, and compactness). These features do not correspond, however, to an overall collapsed polymer state, but rather to the occurrence of a compact (but not ''maximally compact'') core and a swollen exterior.

Let us summarize the conclusions derived from the observations above. Depending on protein length and compactness, we find that the geometrical measures $R$ and $R_G$ (or $R_G^*$) do scale differently with the number of residues for actual protein native states. The *average values* for their scaling exponents $(\overline{\nu_i},\overline{\nu_o})$ appear to be bound differently between the characteristic size exponents for random polymers: $1/3<\overline{\nu_i}<1/2$ for the ''inner size'' and $1/2<\overline{\nu_o}<3/5$ for the ''outer size.'' We find an indication that compact proteins with less than 300 residues appear to have $\nu_i\approx\nu_o\approx 1/3$, whereas longer compact proteins are packed differently ($\nu_i\approx 1/2$ and $\nu_o\approx 3/5$). This work shows that a simultaneous analysis of *different* molecular shape descriptors can provide valuable insights in the case of polymers with constraints in their configurational organization. Whereas one of the descriptors (either $R$ or $R_G$) may be redundant for random homo- or heteropolymers, they do behave differently when the polymer exhibits strong monomer-monomer and monomer-solvent interactions. In other words, one should not assume in general that $R$ and $R_G$ share the same behavior in heteropolymers. The possibility of distinct scaling should be tested in each case.

We should stress that the present $\overline{\nu_i}$ and $\overline{\nu_o}$ exponents must be taken only as indicators of differential size scaling. Their values do not necessarily imply that protein chain configurations are found in the same state as polymer chains in ideal or good solvents. Indeed, other interpretations are possible. The difference in scaling exponents may also represent an effect of side-chain branching on the size scaling of linear backbones. Note that randomly branched polymers belong to a distinct universality class [19,20]. If one views side chains as quenched random branches, a mean-field estimate in the regime of excluded volume would be $\nu_i=1/2$ (instead of $\nu_i=3/5$) [19]. It is possible that our results reflect indirectly the modulation of backbone size caused by the distinct location of hydrophilic and hydrophobic branches throughout the protein.

In closing, we should point out that the distinct size scaling in native states can also provide insights into the structure of folding intermediates. For example, recent experimental results on molten globules [21] indicate a hydrophobic core and a content of secondary structure that can be either substantial [22,23] or rather small [23,24]. It would be valuable to compare $R$ and $R_G$ within a series of compact intermediates corresponding to various proteins [23]. (An approximate $R$ value can be estimated from the mean hydrodynamic radius in solution.) If the lack of secondary structure allows maximal compactness throughout the protein, one would expect $g\approx 1$ for the scaling of molecular sizes. On the other hand, a result $g<1$ for both intermediates and native states will be a strong indication that the differential sizes of protein layers are determined at the onset of hydrophobic collapse, and little affected during the remaining steps of the folding path. A new series of experiments should settle the actual $g$ value.

This work was supported by FRUL (Laurentian University) and by an operating grant from NSERC (Canada).

[1] S. H. White, Annu. Rev. Biophys. Biomol. Struct. **23**, 407 (1994).

[2] C. Tanford, *The Hydrophobic Effect* (Wiley, New York, 1973).

[3] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).

[4] M. Karplus and E. I. Shakhnovich, in *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992), Ch. 4.

[5] T. Garel, H. Orland, and D. Thirumalai, in *New Developments in Theoretical Studies of Proteins*, edited by R. Elber (World Scientific, Singapore, 1994).

[6] T. Garel, L. Leibler, and H. Orland, J. Phys. II (France) **4**, 2139 (1994).

[7] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).

[8] G. A. Arteca, Phys. Rev. E **51**, 2600 (1995).

[9] (a) P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953). (b) P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1985).

[10] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977). (For updated information, see the PDB site at http://www.pdb.bnl.gov).

[11] S. R. Quake, Phys. Rev. E **52**, 1176 (1995).

[12] J.-C. Le Guillou and J. Zinn-Justin, Phys. Rev. B **21**, 3976 (1980).

[13] K. A. Dill, Biochemistry **24**, 1501 (1985).

[14] G. A. Arteca, Phys. Rev. E **49**, 2417 (1994).

[15] P. Biswas and B. J. Cherayil, J. Chem. Phys. **100**, 4665 (1994).

[16] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[17] P. R. Privalov, Adv. Protein Chem. **35**, 1 (1982).

[18] J.-R. Garel, in *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992), Chap. 9.

[19] M. Daoud, P. Pincus, W. H. Stockmayer, and T. Witten, Jr., Macromolecules **16**, 1833 (1983).

[20] A. M. Gutin, A. Yu. Grosberg, and E. I. Shakhnovich, Macromolecules **26**, 1293 (1993).

[21] O. B. Ptitsyn, in *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992), Chap. 6.

[22] A. L. Fink, Annu. Rev. Biophys. Biomol. Struct. **24**, 495 (1995).

[23] C. Radfield, R. A. Smith, and C. M. Dobson, Struct. Biol. **1**, 23 (1994).

[24] V. R. Agashe, M. C. R. Shastry, and J. B. Udgaonkar, Nature **377**, 754 (1995).