ELSEVIER

Journal of Molecular Structure (Theochem) 537 (2001) 319–361

THEO
CHEM

www.elsevier.nl/locate/theochem

Peptide and protein folding☆

G.A. Chasse[a], A.M. Rodriguez[b], M.L. Mak[a], E. Deretey[a], A. Perczel[c], C.P. Sosa[d], R.D. Enriz[b], I.G. Csizmadia[a,*]

[a]*Department of Chemistry, University of Toronto, Lash Miller Chemical Laboratory, 60 George Street, Toronto, Ontario, Canada M5S 3H6*
[b]*Departamento de Química, Universidad Nacional de San Luis, 5700 San Luis, Argentina*
[c]*Department of Organic Chemistry, Loránd Eötvös University, Pázmány Péter Sétány 1/A, H-1117 Budapest, Hungary*
[d]*CRAY Inc., 655 F Lone Oak Drive, Eagan, MN 55121, USA*

Received 1 June 2000; accepted 4 July 2000

Abstract

Ab initio peptide folding, and its role in the reductionistic approach towards the understanding of protein folding are discussed from the points of view of past, present and possible future developments.

It is believed that after the initial holistic approach, we are now at a new epoch, which will be dominated by reductionism. New quantitative mathematical models will be the result of the reductionistic approach that will lead toward a new, more sophisticated holistic era. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Peptide folding; Protein folding; Reductionism and Holism

1. Holistic and reductionistic approaches to protein folding[1]

1.1. Holism and reductionism in scientific research

At the beginning of the twentieth century, Ira Ramsen made the following declaration: [1]

> I always feel like running away when any one begins to talk about proteids in my presence. In my youth I had a desire to attack these dragons, but now I am afraid of them. — They are unsolved problems of chemistry; and let me add, they are likely to remain such for generations to come. — Yet every one who knows anything about chemistry and physiology, knows that these proteids must be understood, before we can hope to have a clear conception of the chemical processes of the human body.

Many things have happened since, yet protein folding remains a formidable challenge, in spite of

☆ Dedicated to Professor Serafín Fraga on the occasion of his 70th birthday.

* Corresponding author. Tel./Fax: +1-416-978-3598.

E-mail addresses: rutile@fixy.org (G.A. Chasse), amrodri@unsl.edu.ar (A.M. Rodriguez), mmak@alchemy.chem.utoronto.ca (M.L. Mak), ederetey@alchemy.chem.utoronto.ca (E. Deretey), perczel@szerves.chem.elte.hu (A. Perczel), cpsosa@cray.com (C.P. Sosa), denriz@unsl.edu.ar (R.D. Enriz), icsizmad@alchemy.chem.utoronto.ca (I.G. Csizmadia).

[1] The term “protein folding” covers both the dynamics (i.e. time dependent) as well as the static (i.e. time independent) aspects of the folding process. The latter aspect deals with the final result: the native 3D structure. The present paper concentrates on the folded 3D structure of peptides and proteins, in other words, the final stage of the folding dynamics.

PII: S0166-1280(00)00687-4

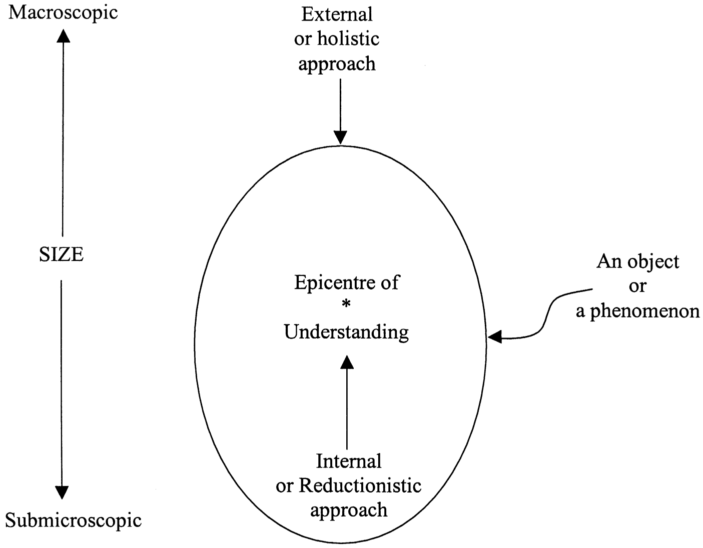


Fig. 1. A schematic representation of the external or holistic and internal or reductionistic approach for understanding an object or a phenomenon.

substantial progress in recent years [2]. Ramsen's conclusion that the full understanding of protein folding has to await "for generations to come" is probably still valid a century later. A complete atomistic mechanism of protein folding most likely will not be revealed in the near future.

It is laudable that experimentalists as well as theoreticians have undertaken the daunting task to identify and explain a detailed mechanism of protein folding process. Clearly, this effort will require not only new hardware, but also a new level of understanding.

For this reason, the term *understanding* needs to be defined. To use an allegorical example, we may compare *understanding* to an onion. There are many layers of understanding, but while we are peeling off layer after layer, there are still many more to be removed. The process of gaining understanding, while painful, is a never-ending story. At this point, we need to discuss the relationship between the *external* or *holistic* approach and the *internal* or *reductionistic* approach to reach understanding. Fig. 1 illustrates schematically the idea of these two approaches.

Let us use the developments of chemistry and biology to illustrate the relationship of the holistic and the reductionistic approaches. Modern chemistry is no more than 400 years old. However, if we consider its predecessor, alchemy, which we may call *pseudochemistry*, then the field is probably over 2000 years old (Fig. 2).

Chemistry started with an external or holistic

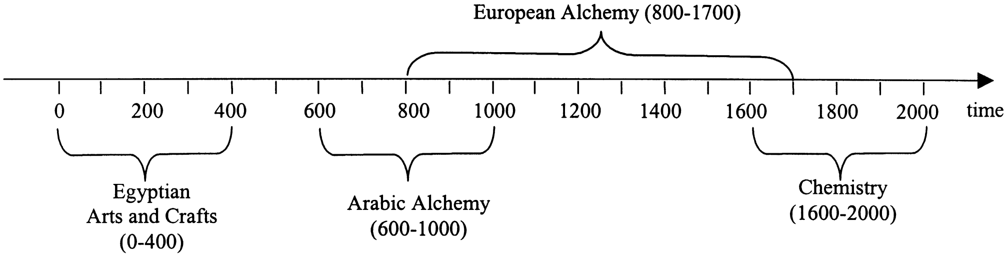


Fig. 2. A schematic time-scale illustration of the development of chemistry and pseudochemistry or alchemy.

approach — materials were regarded as visible objects, they were examined and their properties were studied. The foundation of chemistry perhaps was laid during the development of metallurgy in the sixteenth century by Georg Bauer and Philippus Paracelsus. Robert Boyle published his book, *The Skeptical Chemist* in 1661. Georg Stahl (1660–1735) introduced phlogiston theory, Joseph Priestly (1733–1804) discovered oxygen in 1774 and Antoine Lavoisier (1743–1794) explained combustion. Lavoisier's book, entitled *Elementary Treatise on Chemistry*, was published in 1789, shortly before the French Revolution.

Joseph Proust (1754–1826) introduced the idea of what is known today as the *law of definite proportions* (e.g. $CuCO_3$ *always* contains by mass, 5.3 parts copper, 4 parts oxygen and 1 part carbon), while John Dalton (1766–1844) formulated his principle of what is known today as the *law of multiple proportions* (e.g. carbon has two oxides, for one of them, the carbon to oxygen mass ratio is 1:1.33 and for the other, it is 1:2.66). Dalton postulated the "atomic theory" of chemistry in 1803 without any idea of where his speculation would lead at the advent of the 21st century.

Jons Jacob Berzelius (1779–1848) introduced the symbols of chemical elements and established chemical formulas. He published his monograph in 1806 in Swedish, entitled "Organisk Kemi" which relied on the "vitalistic force" concept based on Torbern Bergman's earlier distinction (1770) of organic and inorganic compounds.

The hypothesis of Amadeo Avogadro (1776–1856) proposed in 1811 led to the mole concept. He postulated that *at the same temperature and pressure, equal volumes of different gases contain the same number of particles.* His hypothesis was not accepted for nearly half a century until Stanislao Cannizzaro, utilizing Avogadro's hypothesis, introduced the concept of the molecular formula in 1858. In the meantime, Wöhler dispelled the myth of the "vitalistic force" by synthesizing oxalic acid, in 1824, through the hydrolysis of dicyane. In 1828, Wöhler synthesized urea by thermally isomerizing ammonium cyanate.

During the period of 1858–1861, Friedrich August Kekule, Archibald Scott Couper and Alexander M. Butlerov introduced, independently, structural formulas in which the concept of valence was incorporated. Finally, in 1872, Dimitri Ivanovich Mendeleev (1834–1907) published the first periodic table of the elements. With this accomplishment, inorganic chemistry became as well organized as organic chemistry. We cannot close this chapter of history without mentioning the introduction of the tetrahedral carbon in 1874 by J.H. Van't Hoff and J.A. Le Bell.

It is clear that from 1600 to 1875, the holistic approach dominated chemistry. However, the introduction of atomic and molecular concepts already foreshadowed the future trend of reductionism.

A new chapter opened, however, at the dawn of the twentieth century. In 1898, J.J. Thomson (1856–1940) discovered the electron. Max Planck (1858–1947) introduced the concept of the *quantum* in 1900. Robert Millikan (1868–1953) measured the charge of the electron in 1909. Rutherford (1871–1937) determined in 1911, that in the atom, a heavy nucleus is surrounded by the very light electrons. In 1913, Niels Bohr (1885–1962) produced the first model of the hydrogen atom, and thereby introduced the early quantum theory. In 1926, the birth of Quantum Mechanics came about with the wave equation of Erwin Schrödinger (1887–1961). Hartree, Fock, Slater, Huckel, Pople, Pariser and Parr appeared on the scene, and with the development of the digital computers in the 1950s, real molecular computations flourished in the laboratory of Prof Robert S. Mulliken (1896–1986) in Chicago. After the early pioneers: C. Roothaan, S. Fraga, B.J. Ransil, E. Clementi, many computational chemists were inspired in Chicago by Mulliken. From this turmoil, it seemed like all hell broke loose, yet nothing like that actually happened. It was only the dawn of another era in the first half of the twentieth century — the reductionistic phase of chemistry started.

Many things took place during the second half of the twentieth century. There were developments in software (e.g. POLYATOM → IBMOL → GAUSSIAN), in hardware (from vacuum tube IBM 709 → transitorized IBM 7090/94 → S360 → CRAY supercomputers, as well as the rapid development of the PC). Initially, most papers were published by the American Physics Society and the American Chemical Society, while specialized journals appeared later on (Theoretica Chimica Acta → Journal of Computational Chemistry → THEOCHEM, etc.). However, these were merely finishing touches to what happened during the first half of the twentieth century.

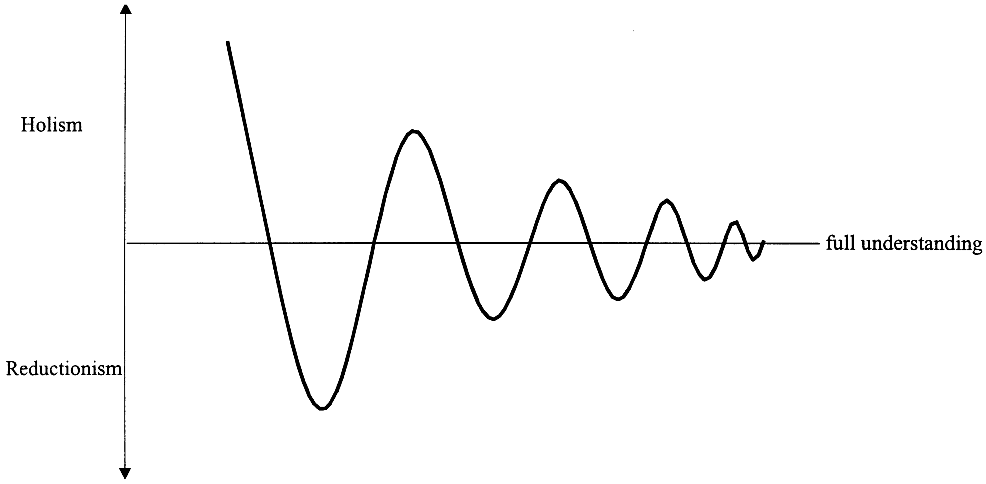


Fig. 3. A schematic illustration that full understanding is reached in an oscillatory fashion with alternating holistic and reductionistic approaches.

At one point, Enrico Clementi declared — "*we can compute anything*". While it is still not literally true for any desired accuracy for any molecular size, and many considered his *eureka* as a poetic expression, nevertheless, he summed up correctly the essence of the new era. Thus, we can see today that reductionism in chemistry is thriving.

In our attempts to understand the relationship between holism and reductionism, let us examine biology very briefly as our second example.

Carl von Linne (1707–1778), or Carolus Linnaeus in latinized form, published his *Genera Plantarium* and *Species Plantarium* in 1737 and 1753, respectively, both works became the basis of modern biology. Without his immense work to classify all known plants and animals, Darwin (1809–1882) could not have developed his theory of natural selection, which was published in 1859 under the title, *On the Origin of Species*. The Austrian born monk Gregor Johann Mendel (1822–1884) published his work on garden peas in 1866. His publication was virtually ignored as a significant work until the early 1900s. The term *genetics* was not coined until 1906 by the British biologist William Bateson. Almost at the same time, nucleic acids were first isolated from fish sperm, but their great biological importance did not become recognized until Francis Crick and James Dewey Watson published the DNA structure during the period of 1951–1953.

Although British biochemist, Sir Hans Adolf Krebs, and Hungarian biochemist, Albert Szent-Györgyi von Nagyrapolt (1893–1986) did advance our understanding of the molecular basis of life processes during the 1930s, through the discovery of the Krebs/Szent-Györgyi cycle or the citric acid cycle, molecular and structural biology did not become recognized fields until the second half of the twentieth century.

Thus, biology was holistic from the mid eighteenth to the mid-twentieth century. However, during the second half of the twentieth century, it gradually became reductionistic. Molecular computations are now being done on biologically important or bioactive compounds such as drugs, and mechanisms of actions are assessed at the electronic level. Molecular recognition, which may well be the basis of immunology, is one of the hot topics debated.

All of these trends are of course, related to medicine. The American Chemical Society publishes the *Journal of Medicinal Chemistry*, Elsevier publishes under Current Trends: *Molecular Medicine Today*. THEOCHEM publishes special issues under the title of *Computational Medicinal Chemistry* and papers are written in the area of *Prospects in Computational Molecular Medicine*[3]. A general practitioner may find such a situation bewildering and may wonder where all of these will lead. Is he not supposed to heal the whole body rather than just one of its molecular components? He surely is right in his

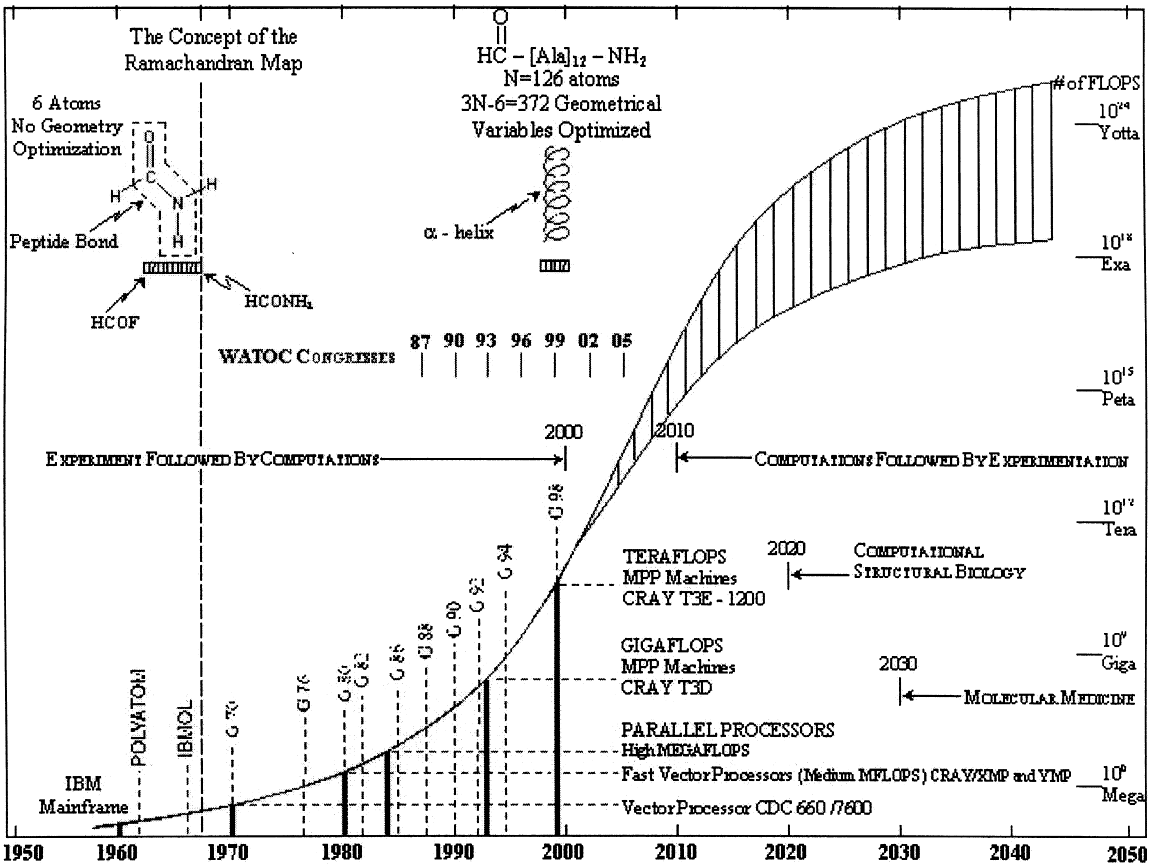


Fig. 4. Development of computer hardware and software as the basis of biomolecular computations. Note that after this figure has been completed (July 1999), IBM announced the creation of the "Blue Gene" ultra supercomputer with one petaflops capability, to be ready by 2005. This makes the optimistic (upper) growth curve a conservative estimate of the future development.

questioning because the human body is more than just a collection of molecules even if it is true that:

> Every disease starts at the molecular level.
>
> Thus, ultimately every cure has to be achieved at the molecular level.

So what does a medical doctor need? Holism or reductionism? As a matter of fact, he needs both because the two approaches are not antagonistic, but complementary. He must realize that there is a phase transition — holism occurs first, followed by reductionism. However, after reductionism, we have more refined holism, in an oscillatory fashion, slowly converging towards full understanding as shown in Fig. 3.

To use another allegoric example to illustrate understanding, we may use the climbing of stairs. We may start with the right foot, but after that we need to step further with the left foot, and then the right foot again. If the right foot is holism and the left foot is reductionism, then it is clear that we need both of them, one after the other, to climb up the stairs of knowledge.

In this paper, we wish to demonstrate that to understand protein folding, we have started with the holistic approach just like any other scientific inquiry. However, we also need to understand the conformational intricacies of single amino acids, as well as the folding of short peptides that is built upon such studies with the reductionistic approach. After gaining sufficient insight to the smaller

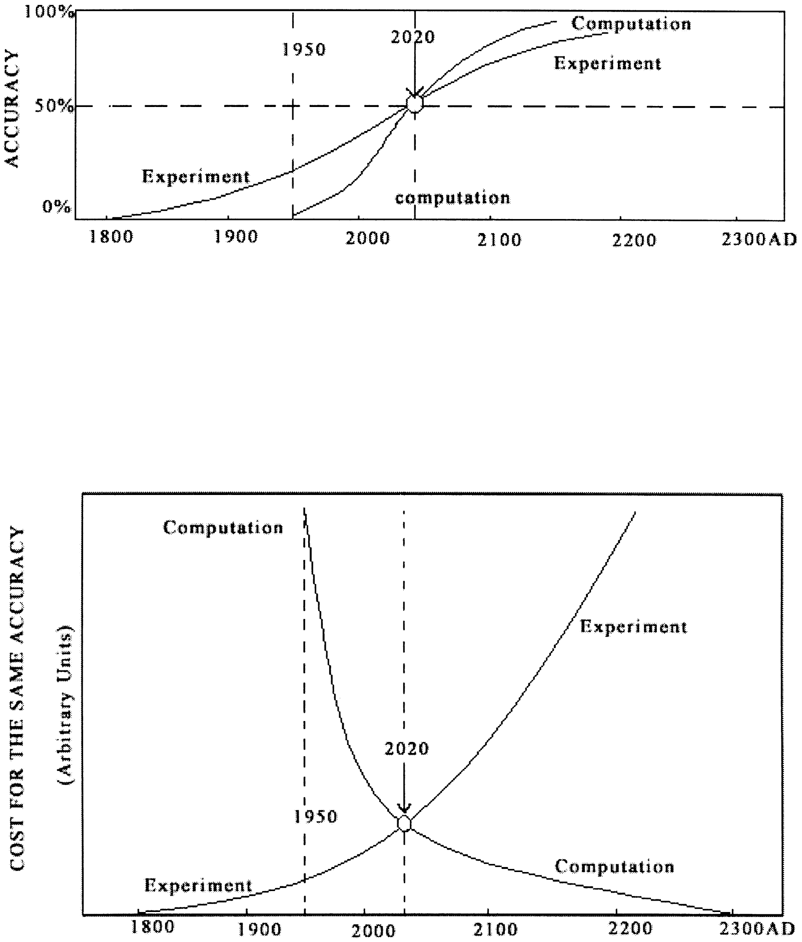


Fig. 5. Predicted cross-over of computations and experiment.

building blocks, we may hopefully be able to return to the holistic approach with considerably more new knowledge, new technical skills, as well as more sophisticated research tools.

1.2. When does the paradigm shift occur from holism to reductionism?

In some sense, every demarcation date is arbitrary. On what date did the medieval times end? Did it actually happen when Columbus landed in the New World or do we just assign an unusual event to the end-date of an epoch? A paradigm shift does not occur in such a way that in one morning we wake up and everything that was important yesterday becomes meaningless and those things we have ignored previously suddenly come to the centre of attraction. There must be a realistic basis for a paradigm shift to take place. In chemistry, holism ended and reductionism started perhaps with the discovery of the electron (1898) or with the introduction of the concept of quantum (1900). This prompted the subsequent development of quantum mechanics (1926). In biology, the corresponding paradigm shift might be assigned to the Watson–Crick publications of the DNA structure and its subsequent consequences (1951–1953).

In protein folding, the newly developed computer technology and necessary softwares made the dramatic change right around the dawn of the third millennium (say, the year 2000, for the sake of simplicity). Now is the time for a changeover from holism to reductionism in the field of protein folding.

In the 1960s, up to four or six atoms could be handled at the ab initio level of theory (for examples, molecules such as HCOF [4–6] or the isoelectronic $HCONH_2$ [7,8], the latter structure containing one peptide bond). However, all calculations were performed on a fixed structure, without geometry optimization using uncontracted basis sets. By the turn of the millennium, the largest peptide on which ab initio calculations have been carried out, contains 12 amino acids [9].

Fig. 4 shows the century of development [3], from 1950 to 2050. In 1950, Boys suggested the use of Gaussian-type orbitals and Roothaan published his SCF procedure in 1951. However, in the 1950s and early 1960s, all digital computers were vacuum-tube based. The first Gaussian calculation on an organic molecule (HCOF) was carried out in 1963 on an IBM 709 computer that was still using vacuum tubes. The leading software was POLYATOM at that time. The first transistorized mainframe computer (such as the IBM 7090 and 7094) arrived a bit later on the scene.

Hardware development after transistorization, in units of FLOPS (Floating point Operations Per Second) is shown by the vertical bars in Fig. 4. Clearly, there was a dramatic change from mega (10^6) FLOPS through giga (10^9) FLOPS, to the current tera (10^{12}) FLOPS. However, it has been suggested that future computers can solve problems in 30 s — what today's 10^{12}-FLOPS supercomputers would take 10 billion years to solve. The ratio (10^{10} years: 30 s) is of the order of 10^{16}. Thus, we may consider the limit of the growth-curve to be $10^{12} \times 10^{16} = 10^{28}$ FLOPS. Fig. 4 shows two curves, the lower of which is the pessimistic prediction and the upper of which is the optimistic one. Note that even the optimistic curve levels out at about 10^{25} FLOPS, which is about 1000 times more conservative than the prediction, which is currently lingering around 10^{28} FLOPS.

This dramatic hardware development is paralleled to a heroic effort of software development. Such development is not quantified by values of benchmark computations; they are simply presented at their times of appearances. The most durable package is GAUSSIAN [10], which has had several editions from 1970 to 1998. The dates of the WATOC congresses (from 1987 through 2005) where molecular computational chemists and biologists report their progress are also indicated. Finally, the dates of the first, and therefore the smallest [$HCONH_2$] [7,8], and last, and therefore largest [$HCO\text{-}(Ala)_{12}\text{-}NH_2$] [9] ab initio peptide calculations of the twentieth century are also marked in.

A great deal has happened during the past 40 years and there is more to come. Currently, we can predict enthalpies of reaction within 1–2 kcal/mol when compared with experiments. Thus, today, experimentation provides us with the primary standard; however, this may change in the future. Computed intrinsic molecular properties, such as energy, may in fact be the primary standard. One may estimate that by the year 2020, computational accuracy will supersede experimental accuracy and it will be cheaper to compute than to carry out experiments (Fig. 5). As far as molecular geometry is concerned, we can compute C^{α}–H bond lengths of glycine more accurately than they can be measured by NMR or by neutron diffraction. Our older colleagues do not believe that this will happen so soon; our younger colleagues are convinced that we shall reach that day sooner. So, let us set the date to be with some tolerance: 2020 ± 5.

It is believed that such dramatic developments as shown in Figs. 4 and 5 are in fact the basis of a paradigm shift, which results in a change of method from holistic to reductionistic in our search for understanding the secrets of protein folding.

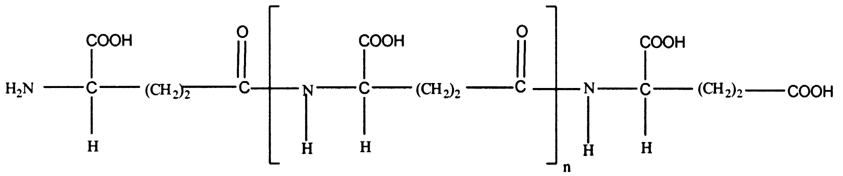
2. The field of macromolecular conformations

2.1. Proteins

Most macromolecules are polymers of smaller units, which are frequently referred to as monomers. This is true for natural, as well as manmade polymers. The former category includes DNA, polysaccharides and proteins, and the latter case consists of materials such as polyamides (nylons), polyesters and vinyl polymers. Obviously, the molecular composition of the polymer is closely related to the physical and chemical properties of these macromolecules. For example, a vinyl polymer will never behave like a protein and vice versa. However, within their structural limitations, the physical and chemical properties of polymers may vary with their conformations. This is true for proteins as well as for vinyl polymers. In the former case, we may notice that egg white and silk are completely different. This is because globular proteins, such as egg white, which contain a lot of α-helical structural motifs, and hence are largely water soluble, while extended β-sheets, as in the tertiary structure of silk, are not. Such structural differences can have biomedical importance, for example, α-helix to β-sheet conversion (i.e. from soluble to insoluble proteins) is related in prion-caused diseases [11]. The conformational dependence of physical and chemical properties of manmade polymers may be less spectacular; nevertheless, its economical and social consequences are enormous.

Many of the macromolecules are linear polymers, but some involve branching. For instance, some naturally occurring polysaccharides (e.g. amylopectin and glycogen) show heavy branching. Manmade polymers, such as dendrites, are designed to have extensive branching.

There are also copolymers, which are based on more than one type of monomer. Most proteins are copolymers of 20 naturally occurring α-amino acids. There are, however, rare exceptions. For example, the cell wall of the anthrax bacteria (*Bacillus anthracis*) is a homopolymer called γ-poly-D-glutamic acid (1):



(1)

This unusual structure was discovered in 1937 by Bruckner and his co-workers [12,13]. It is unusual for three reasons. First of all, it is a homopolymer; secondly, it is not the α-carboxyl but rather the γ-carboxyl group that is involved in peptide bond formation. Finally, it consists of exclusively D-enantiomers, rather than the usual L-isomer of glutamic acid. The molecular structure of this unusual homopolymer has been deduced by degradation [14–19], as well as synthesis [20–24].

Nowadays, manmade polymers are discussed in *materials science*, while natural polymers, such as proteins, fall within the domain of *biomolecular science*. Since these two areas are completely separated, ideas are flowing only rarely from one territory to the other. Protein chemists may be able to learn from polymer scientists and vice versa.

2.2. Manmade polymers

Examination of the complex structures of biopolymers and other macromolecular clusters in solution provides an impetus for continuing research into these aggregates and their interactions with poor solvents. A solvent is designated as ‘poor’ if polymer–solvent interactions become less energetically favourable than the solute–solute ones. The experienced *collapse transition* of a single polymer chain on being introduced into such a solvent is one of the most fundamental phenomena in polymer physics, both because of its academic appeal, and its relation to molecular biology. Issues such as protein folding and their resultant native states, as well as DNA conformations in vivo could be better understood with extensions of current theories and works, in these non-biological areas.

Experiments involving these globules (as they are known) and coil–globule transitions (on being introduced to poor solvents) remain relatively complex, despite the considerable amount of time and empirical efforts spent. Experimental observations with dilute solutions of simple homopolymers, such as polystyrene in cyclohexane, emerge as dubious. It is unclear whether the observed compact particles are indeed equilibrial globules or some sort of aggregates, or other non-equilibrial formations. Alternatively, one may consider the collapse of heteropolymers, where, in the case of natural proteins, heterogeneity can prevent chains from precipitating even in concentrated systems. This cannot therefore serve as a theoretical model for conceptually simpler homopolymer problems.

Hence, homopolymers are studied in depth, both experimentally and theoretically, with two dominant approaches for the latter. The first of these two approaches follows the ideas of Flory [25–27], operating with the RMS polymer size $\langle S_v^2 \rangle$, or using the corresponding ratio with the ideal strain ratio of the radius of gyration.

$$\alpha^2 = \frac{\langle S_v^2 \rangle}{\langle S_v^2 \rangle_0} \tag{2}$$

This is known as an *order parameter*. It describes transition, where v corresponds to the number of segments in the chain [28,30]. The second and more detailed approach makes use of the works of Lifshitz [31,32], employing spatial density distributions of monomers around the polymer centre of mass as the order parameter. This, as in all other studies of this type, makes use of a technical simplification in addition to the fundamental approximations. This simplification is based on the truncation of the virial expansion, in order to retain theoretical control through the inclusion of interactions between monomers. In doing so, the second virial coefficient is usually assumed to represent the attractive part of interactions. The third virial term represents repulsion, which formally prevents the unphysical situation of complete chain collapse to a point (i.e. it provides the mathematical basis for the limiting of density of the globules/aggregates formed). Without integration of this third term, the theory cannot lead to valid conclusions on high density polymer-solutions, due to either $T \ll \Theta$ (see below) or high concentration.

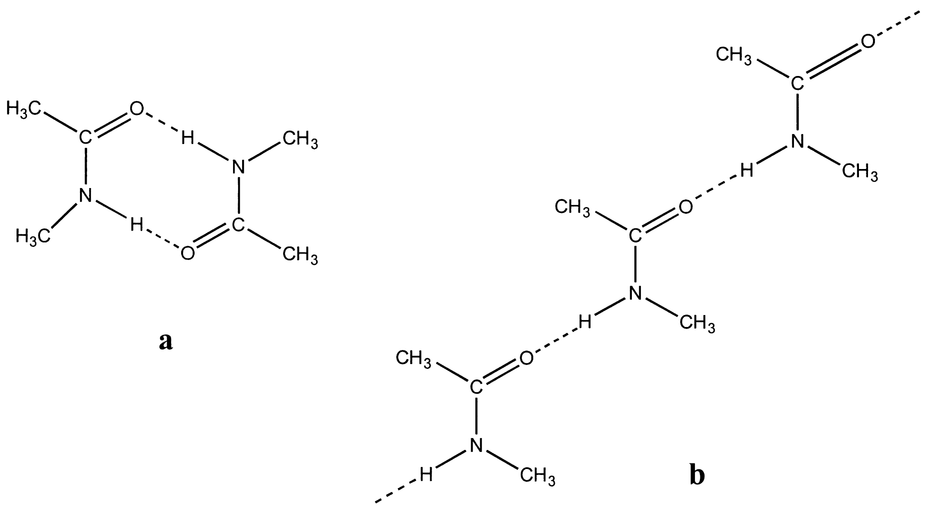
The classification of solvents is temperature-dependent and we find poor solvent characteristics at $T < \Theta$, where Θ is Flory's ideal temperature. Θ can be understood as the point where the second virial coefficient goes to zero; more specifically where solute–solute interactions are absent. One must keep in mind that in some cases, $T > \Theta$ can also bring about a poor solvent condition for certain polymers, containing T-sensitive inter-chain interactions (like hydrogen bonding). On raising T, these stabilizing interactions could be overcome, providing the environment for more favourable associations between like-chain segments (i.e. aggregation).

The process of collapse from coil or non-globular form to an aggregated state, is dominated by long-range, two-body interactions which become attractive in poor solvents. It is also restricted by configurational entropy (this is the intra-molecular force which works against chain contraction in low undercoolings), confinement entropy (this is inter-molecular, with respect to the nearest neighbours) and by repulsive interactions (providing resistance to globular formations). These are dealt with using the SCFE (self-consistent free-energy) [29] minimization within a Gaussian-cluster approximation, which is controlled primarily by a balance of the aforementioned short-ranged two-body attractions and three-body repulsions. The analysis of the collapse transition thereby becomes a key point in describing the driving forces for transitions to globular states of homopolymers. Through monitoring of the strain-ratio of the radius of gyration, one can understand the enthalpic (thermodynamic component) and entropic (kinetic component) contributions to this condensed state. This contraction may be discontinuous (first-order), or smooth and continuous (second-order).

Intermolecular interactions and phase behaviour also come into play as important factors. These factors affect the modelling of macromolecular dynamics, and are much more complex, requiring more in-depth studies and computations of greater orders of magnitude. Such an investment into more extensive theoretical studies would give a profound insight into these interactions, which would either limit or propagate chain collapse. Interactions, such as chain entanglement, must eventually form by two chains subsequent to their aggregation, introducing an additional parameter that includes the 'knotting time' associated with this interaction.

Conclusively, theoretical studies find that for the most part, collapse takes place within and in an infinitely small temperature range, where $T^* \rightarrow \Theta$, hence is an enthalpically driven process, as expected and concluded by experiment.

Given both the experimental and theoretical situations, computer simulations seem to be a good choice for furthering such studies. However, a sound basis of algorithms must be constructed from observed empirical trends and even from ab initio treatments of small template representations of the larger systems studied.



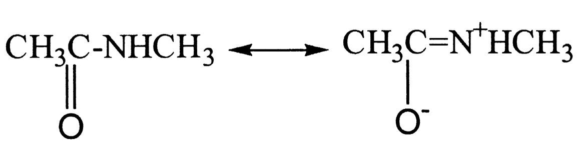
Scheme 1. Possible ways of intramolecular hydrogen bonding of *N*-methylacetamide. a) cyclic, b) linear.

3. Structure determination of proteins

3.1. Early historical background (1900–1950)

One of the simplest molecules which contains a peptide bond, and thus may be considered as a fragment of a polypeptide chain is *N*-methylacetamide, $CH_3CONHCH_3$. The structure determination of this molecule was, therefore, the first problem attacked in relation to the configuration of a polypeptide chain. A number of optical and dielectric data of this substance have been measured by Mizushima et al. [33].

Let us assume from the electronic structure of this molecule that there is resonance (3) for the normal state of this molecule.



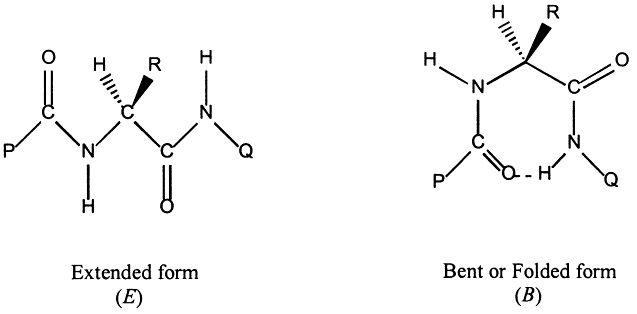
(3)

The resonance energy (see equation (3)) is estimated to be 16 kcal/mol. This is to be compared with the 3 kcal/mol barrier height for rotation about a single C–C bond and about 60 kcal/mol for that of a C=C double bond. Thus, we can see that the peptidic N–C bond has considerable double bond character. Accordingly, only the *trans* and *cis* forms are favourable. Furthermore, it was observed that the hydrogen atom of the NH group tends to form a strong hydrogen bond, and by assuming similar resonance for the polypeptide chain, one could explain the strong tendency of a polypeptide chain to form *intramolecular* and *intermolecular* hydrogen bonds.

The results of Raman and infrared spectra, as well as dipole measurements supported the view that almost all of the molecules are in their *trans* configuration. The concentration-dependence of the molecular polarization observed in carbon tetrachloride indicates that such molecules readily associate in such solutions. If these molecules were in the *cis* form, they would dimerize (Scheme 1a) and the apparent moment would become larger with decreasing concentration.

This is not compatible with the results of dipole measurements in which the apparent moment has been found to decrease from 6.6 to 4.8 D with decreasing concentration. If, however, the single molecule has a *trans* configuration as suggested by Raman and infrared measurements, the associated molecule will have a chain configuration where the dipole moment will be larger than that of a single molecule. This would, therefore, account for the concentration-dependence of the apparent moment quite well (Scheme 1b) [34–37].

It is worthy to note that the main chain of polypeptide showed near-infrared absorption that is quite similar to



Scheme 2. Extended bent forms of a single amino acid diamide.

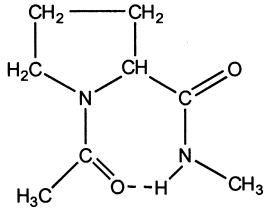
that of the associated *N*-methylacetamide. Accordingly, it was suggested that the peptide bond in the polypeptide chain could also be in the planar *trans* form.

The peptide bond also keeps the planar *trans* form in molecules with two peptide bonds P–CONH–CHR–CONH–Q, although these molecules can exist in different configurations. Molecules with two peptides bonds can take both the extended (E), and folded or bent (B) forms, with stable positions of internal rotation potential, including that corresponding to the planar configuration of the peptide bond (Scheme 2).

In the middle of the twentieth century, it was thought that the extended (E) configuration forms only *intermolecular* hydrogen bonds, but not *intramolecular* hydrogen bonds. The existence of these two forms, according to the measurement of infrared absorption, was proven by Mizushima et al. [38].

The equilibrium ratio of the extended and folded forms is different from one substance to another. In the extreme case of acetylproline-*N*-methylamide in carbon tetrachloride solution, virtually all molecules exist in the folded form since this configuration maximizes *intramolecular* hydrogen bonding opportunities as shown in Scheme 3.

The conclusions drawn above for the structure of molecules with two peptide bonds were useful in constructing a stable model of a polypeptide chain. The tendency to take on an extended form (E) or a folded form (B) for acetylamino acid *N*-methylamide is closely related to the tendency of taking one of these forms for the corresponding amino acid residue in a polypeptide chain. Accordingly, proline would tend to take the folded form.



Scheme 3. Bent form of proline.

At this point of the review of the classical analysis of the structures of molecules with peptide bonds, it would be appropriate to discuss the strength of hydrogen bonds to some extent.

When a group, X–H, is involved in hydrogen bonding (X–H···Y), the X–H absorption is shifted to lower frequencies, as in the case of the N–H···O system of *N*-methylacetamide, acetylglycine *N*-methylamide, and so on. An estimation of the strength of hydrogen bonds can be made from the amount of the shifts in frequency, as has been suggested by several investigators [39,40]. The hydrogen bond between N–H and O=C groups of two peptide bonds is much stronger than that formed between two similar but separate groups of, for example, acetone and alkylamine. In the case of *N*-methylacetamide, this is due to resonance (3), according to which the proton accepting power of the amide CO group becomes stronger than that of a ketone. Additionally, the amide proton becomes more acidic than the hydrogen of alkylamine. Thus, the contribution of the polar structure becomes greater, once the peptide bond is involved in hydrogen bonding of N–H···O=C. Accordingly, the N–H···O hydrogen bonding in associated molecules shown in Scheme 1b becomes stronger, as the chain of the associated molecules becomes longer [41,42]. The situation will be similar in the case of

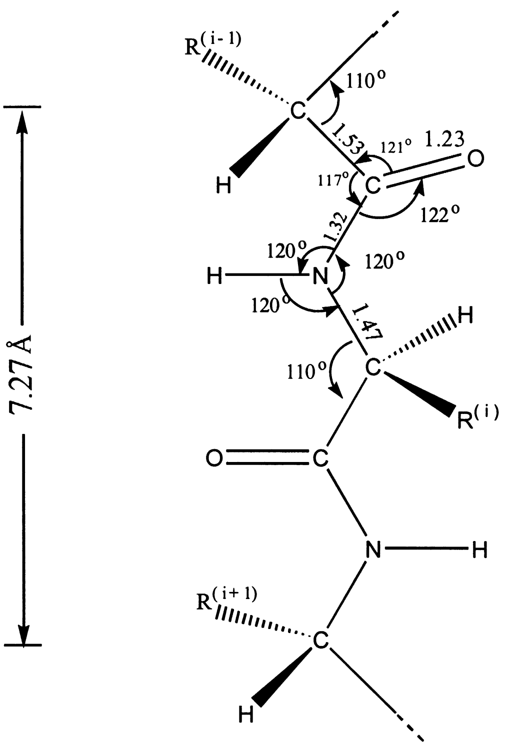


Fig. 6. Structural representation of an extended peptide chain in the mid twentieth century.

molecules with two or more, peptide bonds. Therefore, the intramolecular and the intermolecular hydrogen bonds of similar type will be formed fairly strong within or between polypeptide chains.

Amino acids are also simple substances closely related to proteins. Albrecht and Corey [43] determined crystallographically the configuration of glycine. The result of crystal structure determination by the same investigators suggests strongly that the crystalline molecule exists as a zwitterion with positively charged amino group and negatively charged carboxyl oxygen atom. The electrostatic force due to these positive and negative extremes as well as the hydrogen bonds between nitrogen and oxygen atoms firmly folds the molecules in the lattice.

The zwitterionic character in crystalline glycine was also suggested by Raman measurement by Baba et al. [44]. These investigators also measured the Raman spectra of glycine aqueous solutions and found the structures, $^{+}NH_3$–CH_2–COO^{-}, $^{+}NH_3$–CH_2–COOH and NH_2–CH_2–COO^{-} in the neutral, acidic and basic solutions, respectively. Their results in solutions are in agreement with those of the earlier investigators, among whom Edsall and his coworkers [45–50] made extensive measurements not only on glycine but also on other amino acids and related compounds.

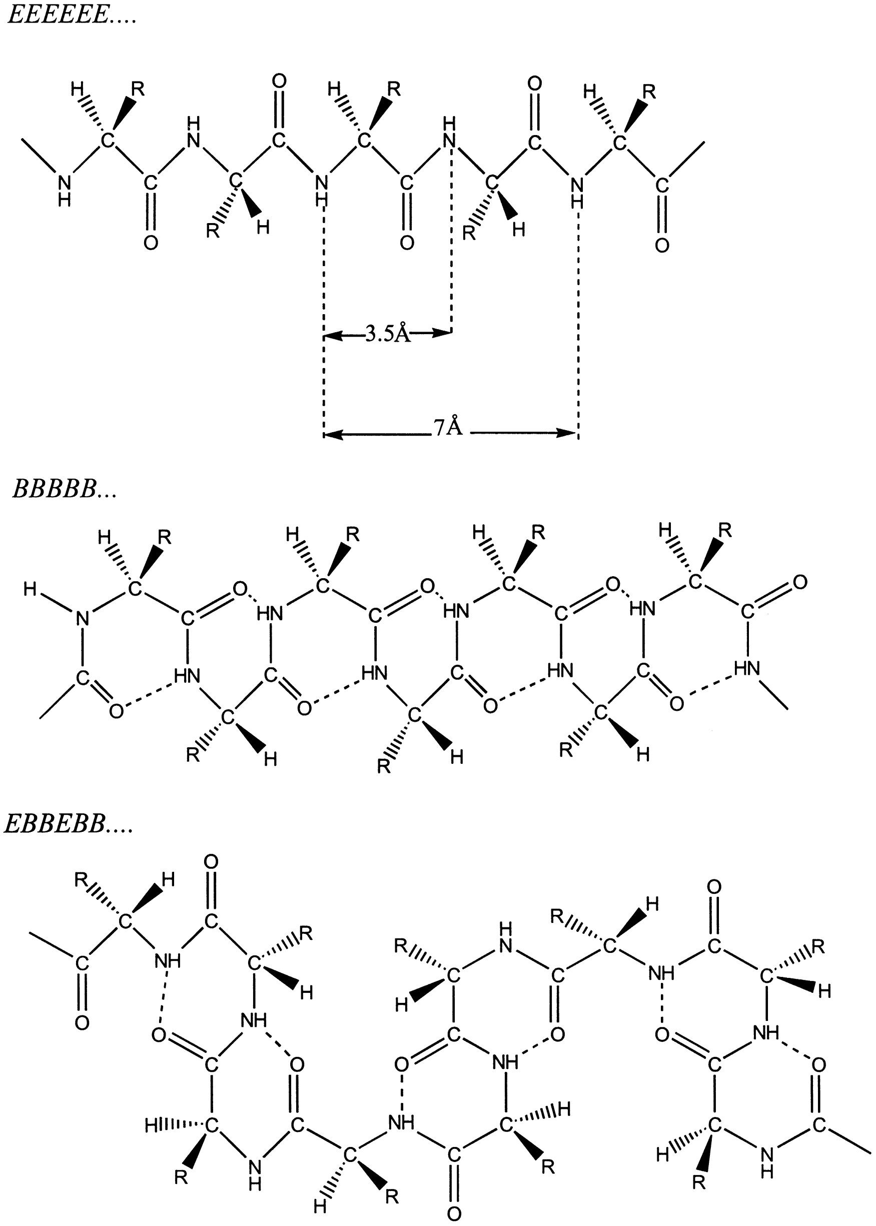
In infrared absorption, the evidence for the zwitterionic structure of glycine and other amino acids is equally strong [51,52]. An interesting fact first noticed by Wright [53,54] is that the infrared spectrum of the D,L-form of an amino acid is usually different from the spectrum of either the D- or the L-form of the same acid when each is examined in the solid state. Darmon et al. [55] confirmed this observation.

X-ray crystal structure studies have also been performed on other amino acids. The results gave us important information concerning the lengths of covalent bonds between carbon, oxygen and nitrogen atoms and the angles that these bonds make with one another. Fig. 6 represents the probable dimensions of fully extended polypeptide chains as suggested by the paper of Corey and Donohue [56].

Due to the internal rotation about a single bond, a polypeptide chain can assume various configurations, of which the simplest one is the fully extended (E) configuration denoted (EEEEEE…) (Fig. 7). Meyer and Mark [57] were the first to show that silk threads contain such extended chains in which the length of a peptide unit in the chain amounted to 3.5 Å.

Some molecular models of α-keratin have been proposed by Astbury [58,59], Huggins [60] as well as Shimanouchi and Mizushima [61,62]. The last mentioned investigators proposed the fully folded configuration denoted (BBBBBB…), as well as the extended and folded combined form (EBBEBB…) (Fig. 7). Originally, these two structures were proposed by Shimanouchi and Mizushima [61,62] on the basis of the internal rotation potential including the *trans* planar configuration of the peptide bond. In other words, all the movable atoms or groups in the main chain are in the stable positions of internal rotation potential.

The folded configuration proposed by Shimanouchi and Mizushima, as well as the extended configuration proposed by Meyer and Mark, satisfy the condition that all amino acid residues contained in the main chain are in L-forms. There are many other



EEEEEE....

BBBBB...

EBBEBB....

Fig. 7. Schematic representations of typical peptide structures in the mid twentieth century (E = Extended, B = Bent or folded).

configurations, which satisfy this requirement and that of the internal rotation potential stated above.

Let us now explain the prevalent level of understanding of such phenomena in about 1950, which is based on various experimental results obtained for keratin and some other proteins of these polypeptide configurations:

1. As already stated, the X-ray diffraction pattern of α-keratin can be explained by BBBBBB...

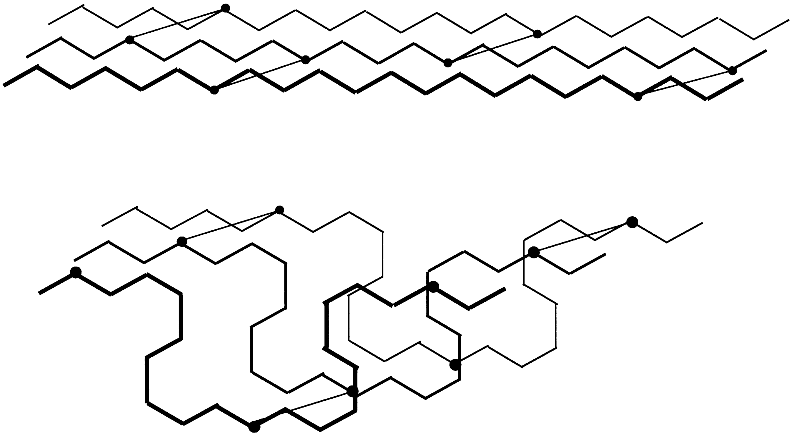


Fig. 8. A schematic representation of the two (α and β) forms of keratin (top: EBBEBB; bottom: EEEEEE).

configuration (fibre period = 5.14 Å) or by BBEBBE… configuration (fibre period = 10.3 Å).

2. Upon α → β transformation, the side chain spacing remains unchanged. This can be seen quite clearly from the configurational change from BBB… to EEE… or from EBBEBB… to EEEEEE…as shown in Fig. 8. Such a change does not affect the cystine bonds connecting the polypeptide chains laterally.
3. The backbone spacing (4.5 Å) of β-keratin does not exist in α-keratin. This is quite understandable because the *intermolecular* hydrogen bonds that keep the neighbouring polypeptide chains together, at the backbone spacing, disappear and form *intramolecular* hydrogen bonds.
4. The supercontraction of wool can be explained if the keratin molecule which was originally in BBBBBB… configuration is changed to take EBBEBB… configuration by the disconnection of the disulfide bridges of cystine units.
5. Early X-ray investigations show that the crystal structures of proteins such as insulin [63] and excelsin [64] have trigonal symmetry. This is understandable if we consider that the molecule of such proteins is made of the ring configuration with trigonal symmetry, or of the suitable superposition of these planar configurations. (The superposition may be caused through hydrogen bonds or through covalent bonds of side chains.)
6. Whether a residue takes the extended or folded form in a polypeptide chain depends upon the nature of an amino acid residue. We have already seen that different molecules of the type, CH_3–CONH–CHR–CONH–CH_3 have different tendencies to assume the extended (E) form or the folded (B) form. Since the configurations of amino acid residues in polypeptides are comparable to those of acetylamino acid *N*-methylamides referred to above, proline residue will show strong tendency to take the folded form. The characteristic combination of the extended and folded forms of different amino acid residues in a polypeptide chain will correspond to the specificity of the protein. The importance of the order of amino acid residues in a given protein will be closely related to the tendency of each residue in taking the extended or the folded form.
7. According to experimental results obtained for acetylamino acid *N*-methylamide, the two *unit configurations*, E and B, do not differ much from each other in their internal energy. In addition, the potential barrier to internal rotation over which E and B can pass into each other is not high. Consequently, if denaturation involves a process of opening the polypeptide chain (i. e. the transformation from B to E), we can understand why some proteins are denatured easily. We have already stated that the entropy change of acetyl amino acid *N*-methylamides on passing from the folded form into the extended form can explain reasonably the entropy change of some proteins on denaturation.

It is interesting to note that a folded configuration, exactly the same as BBB… proposed by Shimanouchi and Mizushima, was also presented by Zahn [65] and by Ambrose and Hanby [66,67] independently. The last-mentioned investigators made measurements with polarized infrared radiation of oriented films on poly-γ-methyl-L-glutamate cast from solution in *m*-cresol and concluded that the polypeptide chains take the BBB… configuration. Bamford and Hanby [68] also reported that some synthetic polypeptides in the BBB… configuration with hydrocarbon side chains are soluble in nonpolar liquids, and that after conversion to the EEE… configuration, the polypeptides become completely insoluble in nonpolar solvent. This is quite understandable, since in BBB…, all the hydrogen bonds are intramolecular, just as in the case of acetylproline *N*-methylamide in carbon tetrachloride. Ambrose and Elliot extended their infrared study to proteins. They showed that in oriented films of β-keratin (swan feather), the N–H bond is predominantly perpendicular to the direction of extension of the polypeptide chain. This is in agreement with the picture of the fully extended chain proposed for these proteins by Astbury and Street [69].

Using the values of atomic distances and bond angles shown in Fig. 6, and assuming that the amino acid residues are equivalent, Pauling and Corey [70] constructed two hydrogen-bonded helical configurations for the polypeptide chain. In these configurations, the peptide bonds have planar structures as in the case of EEE…, BBB…, and so on. Moreover, all NH and CO groups are involved in hydrogen bonding, in which the nitrogen-oxygen distance is 2.72 Å and the vector from the nitrogen atom to the hydrogen-bonded oxygen atom lies no more than 30° from the N–H direction. For a rotational angle of 180°, the *helical* configurations may degenerate to a simple chain with all of the principal atoms, C, N and O in the same plane.

Of these two helices, the 3.7-residue helix (α-helix) is interesting in view of the experimental result obtained by Perutz [71]. He suggested that some proteins or synthetic polypeptides may have this helical structure. However, one must not think that only such helical structures constitute the important part of protein structures. At any rate, we cannot explain the most interesting property of proteins, namely their specificity, by such a uniform structure as they occur in keratin.

During the first half of the twentieth century, scientists were happy to be able to deduce the conformational structure of the highly ordered keratin. Of course, the keratin structure is a far cry from the three-dimensional (3D)-structure of folded proteins such as enzymes and receptors. For the understanding of this important problem, we have to consider the characteristic combinations beyond the basic conformations such as extended, folded and helical forms [72,73]. We can add, in closing, that today, the traditional E and B conformations may be named β_L and γ_L.

3.2. Modern historical background (1950–2000)

3.2.1. X-ray

Our insight of protein structure and function has been deepened immensely by X-ray crystallography. In the first half of the twentieth century, X-ray crystallography was a tedious process, whereby only small molecules could be handled. The first X-ray diffraction pattern of a globular protein was obtained by Dorothy Crowfoot Hodgkin, along with J.D. Bernal. These protein crystals were extremely difficult and tedious to work with in the early 1930s, due to the lack of complimentary technology.

A significant difference between protein crystals and other crystalline substances is that they are extensively hydrated — typically 40–60% water by volume. The water is to preserve the integrity of the protein structure. Therefore, it came as a surprise when protein structures were solved by crystallography in the late 1950s. In 1959, John Kendrew obtained clear 3D structures of sperm whale myoglobin [74]. Kendrew took up the problem of myoglobin which, being a quarter of the size of haemoglobin, seemed a more hopeful candidate for X-ray study. Kendrew shared the Nobel Prize for Chemistry with Max Perutz in 1962 "for their studies of the structures of globular proteins". They also introduced the heavy atom method and solved the structure of haemoglobin in solution state at 5.5 Å. Shortly thereafter, Perutz solved the structures of human deoxyhaemoglobin and horse methemoglobin [75].

Protein crystallography continued to present

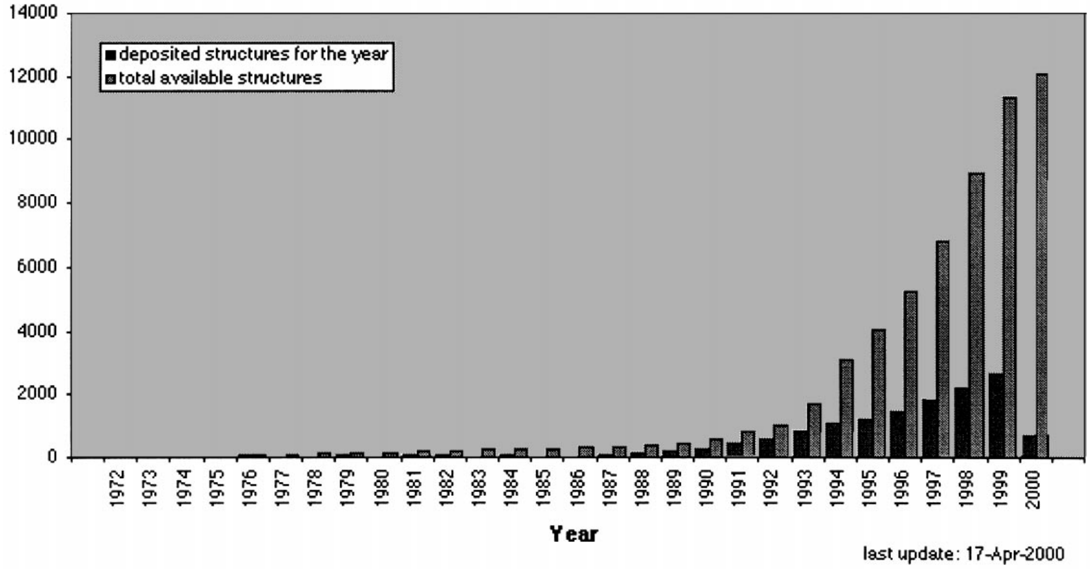


Fig. 9. Number of protein structures deposited in the Brookhaven PDB.

surprises to us throughout the second half of the century.

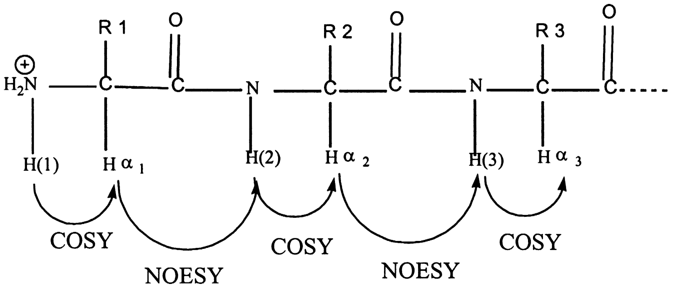
In the 1970s, the surprise was the wide range of structures found, although the majority seemed to cluster into a few general types. In the 1980s, the surprise was that protein crystallography became an indispensable tool of protein chemistry because of the increased speed of determining structures. In the 1990s, many more protein structures were determined. Fig. 9 illustrates the growth of the total number of protein structures deposited in the Brookhaven Protein Data Bank (PDB).

It is clear that the extraordinary development of the field of protein structure determination is in great debt to the past development of protein crystallography. Table 1 illustrates that the PDB holdings not only cover peptides and proteins, but also other biomolecules. It also indicates that in addition to X-ray crystallography, other techniques are also used for biomolecular structure determination.

X-ray crystallography has been recognized as a reliable method since most crystalline proteins take on almost the same structures that they have in solution and therefore maintain their native conformations. There are numerous occasions where different crystal forms of the same protein have been studied, and the molecular conformations are essentially the same. The same can be said regarding solution NMR structures and X-ray structures of the same protein. Most importantly, many enzymes are catalytically active in the crystalline state. Since catalytic activity is prone to the relative orientations of groups that take

Table 1
PDB holding of biomolecular structures as of April 2000 (data taken from http://www.rcsb.org/pdb/holdings.html)

Method	Molecule type				
	Proteins, peptides, and viruses	Protein/nucleic acid complexes	Nucleic acids	Carbohydrates	Total
X-ray diffraction and other	9195	463	513	14	10 225
NMR	1575	65	319	4	1971
Theoretical modelling	243	18	17	0	278
Total	11 013	546	849	18	12 474



Scheme 4. Piston-proton coupling schemes for COSY and NOESY NMR measurements.

part in catalysis and binding, one can safely deduce that crystalline functional enzymes assume conformations that closely resemble the solution conformations.

3.2.2. NMR

The utilization of nuclear magnetic resonance (NMR) spectroscopy to study biologically important molecules began in the 1960s when biochemists realized that the methods used to determine structures of small organic molecules could also be applied in structure determination of much larger biomolecules.

During the past decade, NMR has emerged as a powerful tool for structural studies of proteins. Thus, in addition to X-ray crystallography, there now exists a second method for determination of the 3D structures of small- to medium-sized proteins (up to ~40 kDa). While it cannot compete with X-ray crystallography in terms of the number of structures deposited into the Brookhaven PDB (see Fig. 9 and Table 1), the growth of the number of NMR structures deposited has been very rapid in the past few years.

Although the application of NMR, particularly solution-state NMR, to the determination of the 3D structure of proteins is now well-established [76], the application of two-dimensional (2D) NMR methods to proteins was pioneered only in the early 1980s. Soon after this, Kurt Wüthrich, in collaboration with Richard Ernst, developed to 3D methods. Wüthrich put forward a strategy for carrying out sequential assignments of protein spectra that now forms the basis of all protein-oriented, high-resolution ^{1}H NMR methods. One of the first protein spectra to be completely assigned using this technique was that of a 51-residue fragment of the *Escherichia coli* lac repressor's DNA-binding domain [77].

As an illustrative example, let us consider the proton-based resonance assignment of the peaks in the spectrum of the peptide shown in Scheme 4.

The first step in such a process is to assign the resonances for the protons in the peptide backbone (Scheme 4) using COSY (homonuclear correlation spectroscopy) and NOESY (nuclear Overhauser effect spectroscopy). While the former allows identification of nuclei connected by two to three bonds, the latter can identify nuclei that are in close proximity in space. It is customary to combine COSY and NOESY spectra into a single spectrum because this aids in finding the pattern of alternating COSY–NOESY correlations that extend down the peptide chain. A simulated spectrum for the backbone described in Scheme 4 is shown in Fig. 10. With the backbone protons assigned, it is necessary to continue with the assignment of the side-chain protons. Generally, 20 different side chains occur in proteins (Table 2), corresponding to the 20 natural amino acids, although other, more unusual side chains can also be found in certain situations. The chemical shifts of the proton attached to the alpha carbon of each of the 20 amino acids are shown in Table 3. To carry out this part of the resonance assignment, instead of NOESY, TOCSY (total correlation spectroscopy) spectra are used in combination with the COSY ones. The TOCSY method allows the identification of nuclei within a given spin system and hence it is more useful for side chain assignment than the NOESY technique which correlates nuclei in a way that is independent of the spin system(s) they are in. Schematic COSY and TOCSY spectra for isoleucine are shown in Fig. 11.

Despite its obvious successes, NMR methods for solution structure determination have certain pitfalls. One of the most serious problems is that for larger

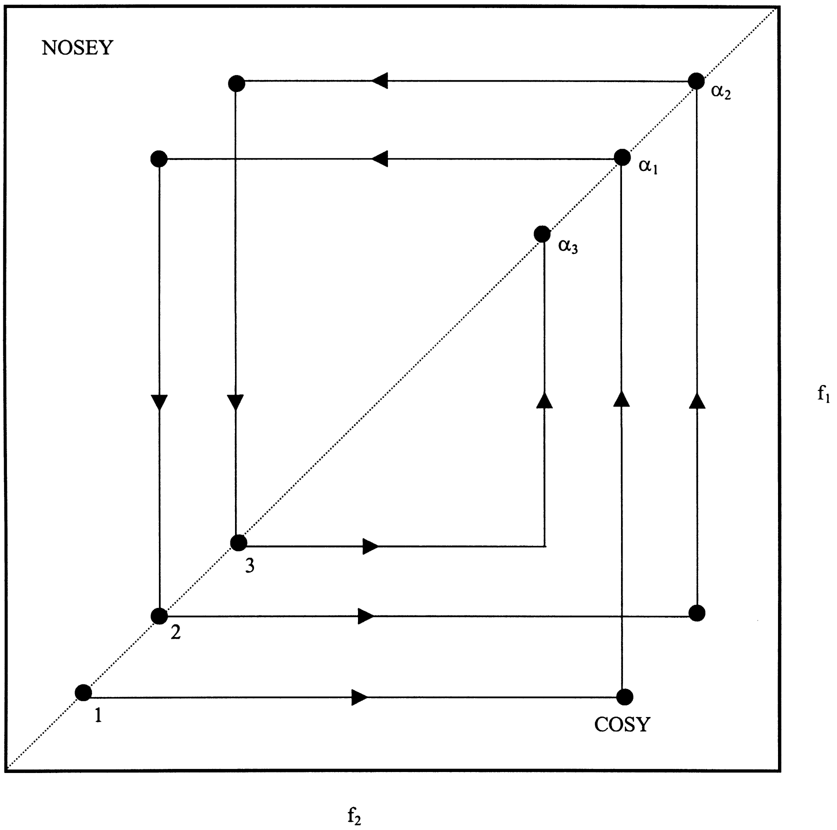


Fig. 10. A schematic representation of the combined COSY/NOESY spectrum of a peptide backbone.

proteins, 2D spectra become very crowded and identification as well as assignment of peaks become difficult, if not completely impossible. A logical way of increasing the resolution is to extend the NMR experiments into third, fourth and higher dimensions; and various reports [78,79] have indicated a promising future for these techniques. A second serious pitfall results from incorrect resonance assignments and/or errors in the evaluation of structural constraints. This will be discussed in Section 3.2.3. Besides structural determination, NMR has also made a major impact in other areas of protein research. For example, NMR studies of protein dynamics are still in their infancy but should, in the future, lead to a new level of understanding, perhaps showing how dynamic properties of proteins affect their biological function(s). Furthermore, 2D NMR spectroscopy can also be used to evaluate the fundamental molecular mechanism of protein folding and new methods have been devised to provide detailed structural insight into the most elusive of entities, the folding intermediate [86–88].

3.2.3. The role of computations in 3D structure determination by NMR

Although there is no “universal” computational method for protein structure determination, there is a common line of attack, as illustrated in Fig. 12. In this general approach, initial structures are generated using distance geometry methods. While such methods rely mainly on geometric constraints obtained from NMR data, general knowledge of the structure’s covalent architecture is also essential. Two different algorithms are used: one operates in distance space (metric matrix approach)

Table 2
Structures of amino acid residues. Note that in addition to the 20 naturally occurring amino acids, which have both DNA and RNA codons, the 21st amino acid, Selenocysteine (Sec), which has only an RNA codon, is also included

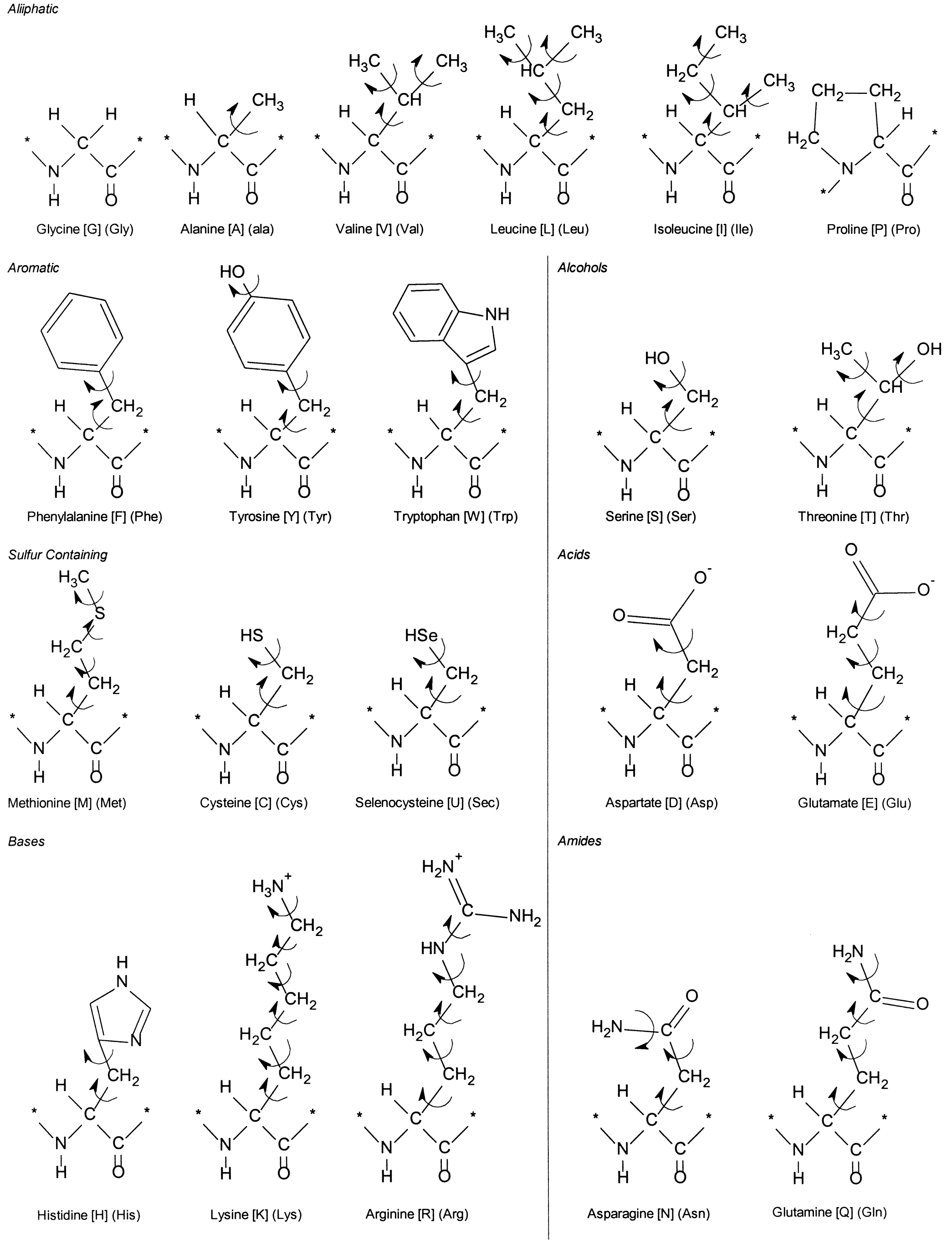


Table 3
Chemical shifts of amino acid protons attached to the α-carbon (data of the residues (X) for non-terminal residues in tetrapeptides GGXA, pH 7.0, 35°C taken from Ref. [76])

Name of amino acid	δ[a]
Glycine	3.52
Alanine	3.76
Valine	3.59
Leucine	3.70
Isoleucine	3.70
Proline[b]	4.11
Phenylalnine	3.98
Tyrosine	4.28
Tryptophan	4.24
Serine	3.85
Threonine	3.51
Methionine	3.83
Cysteine	3.95
Aspartic acid	3.85
Glutamic acid	3.72
Histidine	3.98
Lysine	3.73
Arginine	3.74
Asparagine	3.85
Glutamine	3.73

[a] δ stands for the chemical shifts of amino acid protons attached to the α-carbon.

[b] Data for *trans*-proline.

[80], while the other depends on a variable target function approach with minimization in dihedral angle space [81]. The goal of both approaches is to generate a large number of structures, all of which are consistent with the NMR constraints, by running the programs repeatedly.

3.2.4. Afterthought

NMR spectroscopy is becoming as important as X-ray crystallography in studying small- and medium-sized soluble proteins. A comparison of the solution and crystal structures of bovine pancreatic trypsin inhibitor [82], barley serine protease inhibitor [83] and potato carboxypeptidase inhibitor [83] clearly showed that the global fold established by X-ray crystallography is reproducible in solution by NMR. Since the two techniques independently give the same structures for a number of proteins, it can be concluded that NMR can correctly determine the structure of globular proteins. However, it is important to realize that not all crystal and solution structures are the same. For example, there are substantial differences between the crystallographic and solution structures of rat metallothionein, both in the polypeptide fold and in the coordination within the metal–cysteine cluster [84]. More marked differences have also been found in the global fold of the histidine-containing phosphocarrier protein of *E. coli* [85,89].

4. Statistical predictions of protein folding

New primary structures of proteins are emerging daily. Their discovery naturally leads to our desire to predict their structures in order to understand the structure–function relationships in proteins.

Anfinsen's discovery [90] of the reductive denaturation and oxidative renaturation of ribonuclease led to his suggestion that primary structure specifies the 3D structure of proteins. This triggered interest in protein structure prediction. Many researchers have been instrumental towards the advancement of this field, amongst them is Prof Serafin Fraga. He developed a programme named "MAPSI", or Modelling and Analysis of Protein Structures and Interactions. In addition, his work has been summarized in the book, *Computer Simulations of Protein Structures and Interactions* [91].

One school of thought assumes that a protein folds in order to minimize the free energy of the system. Contributors in this area include Liquori et al. [92], and Ramachandran et al. [93,94], who first showed that a peptide can only adopt certain allowed conformations. The first attempt of predicting secondary structure of polypeptides was performed by Guzzo [95] in 1965. A significant milestone for the prediction of secondary and tertiary structure was laid down by Fasman and Chou [96,97] who developed the first empirical prediction system based on crystallized structures. Its appeal lies in its reliability and ease of use. Prior to the Chou–Fasman rules, there has been a great deal of theoretical efforts; other efforts followed have been reviewed [98]. However, it has been shown that minimization schemes have failed to predict chain folding accurately [99,100]. The difficulty in making sufficiently accurate calculations that are mathematically sophisticated and computationally manageable has been the limiting factor in this approach. Nevertheless, such a

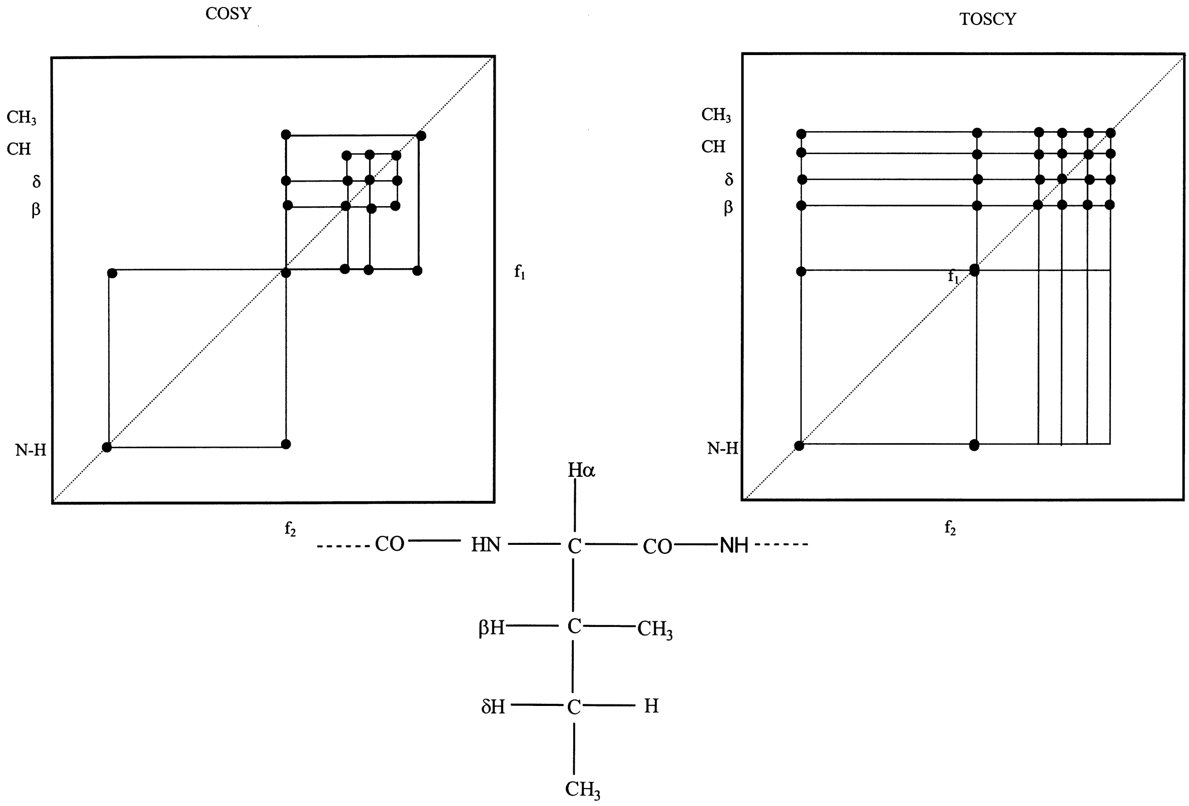


Fig. 11. A schematic representation of the COSY and TOCSY spectra for the isoleucine residue.

theoretical method is the ultimate key to a thorough understanding of how and why proteins fold to their native structures.

The prerequisite for protein prediction is to have the amino acid sequence. Sequence alignment algorithm of Needleman and Wunsch [101] is the basis of one of the earlier attempts to determine whether protein folding was by chance. There are now many algorithms available in the literature. In addition, the local environment of the peptide should also be considered. Hydrophobic interactions and hydrogen bonding are regarded as the two main driving forces of the folding of proteins.

There are many methods to evaluate protein conformations, each of them focussing on a different aspect. A well-known example is the hydrophobicity scale developed by Kyte and Doolittle [102]. The hydrophobic effect has been regarded as pivotal in studies of protein folding, self-assembly and conformation. Solvent accessibility is another factor often studied [103]. It can be treated as the counterpart of hydrophobicity. It provides a quantitative estimate of the probability of an amino acid being buried away from, or exposed to solvent. This probability was found to be proportional to the overall surface area. Other groups, such as Crippen and Kuntz [104], have attempted to relate relative stability to the degree of "ideal packing" by examining the atom density of a central atom, taking into consideration its covalent bonded neighbours and its proximity to the surface of the protein.

There are two main approaches to the prediction of secondary structures. One is based on parameters obtained from the analysis of already determined sequences and structures, and another is based on stereochemical parameters. The Chou–Fasman parameters [96] belong to the former category and have been widely used, particularly for predictions of β

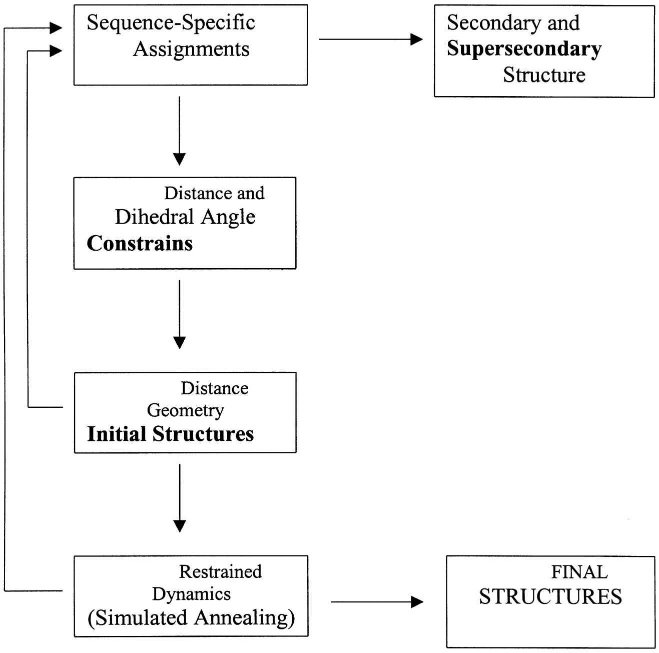


Fig. 12. A flow chart of the structure determination process for a protein.

turns [97]. This is a significant fact since it has been proposed earlier that such turns are possible sites where protein folding originates. It should be noted that care must be taken in the interpretation of results obtained by the method since the exact accuracy is dependent on one's interpretation of the definition of secondary and tertiary structures.

There are different methodologies that have a slant towards various applications. For example, knowledge-based modelling methods can be used for targets of interest in drug design and protein engineering. This approach compares the protein in question to other proteins of known 3D structures at all levels in the hierarchy of protein organization. Another method is the utilization of neural networks, which predict secondary structures of local sequences on the basis of existing protein structures. It was concluded that local information alone is unlikely to provide better results for non-homologous proteins. Computer-aided designs of complex biomolecules are also used. An example of such expert systems is a collection of programmes called LUCIFER, developed by Robson et al. [105]. The method aims to carry out minimizations at the greatest possible speed by judicious use of external data as an aid.

Although the prediction of tertiary structures is at its preliminary stage, one can easily foresee the wealth of knowledge and understanding one can generate in this area. A new Renaissance is expected to follow when ab initio molecular computations will lead to quantitative results of backbone/side chain and side chain/side chain interactions.

The present situation is particularly acute because the sequencing of primary structures is proceeding faster than X-ray determination of secondary and tertiary structures. Thus, the gap between known sequences and known 3D structures is on the rise, as illustrated schematically in Fig. 13. It is natural, therefore, to look for ways and means to predict the 3D-structure from the primary sequences, even if it could only be achieved with partial success.

5. External or holistic modelling of protein folding

Computational modelling of folded protein structures

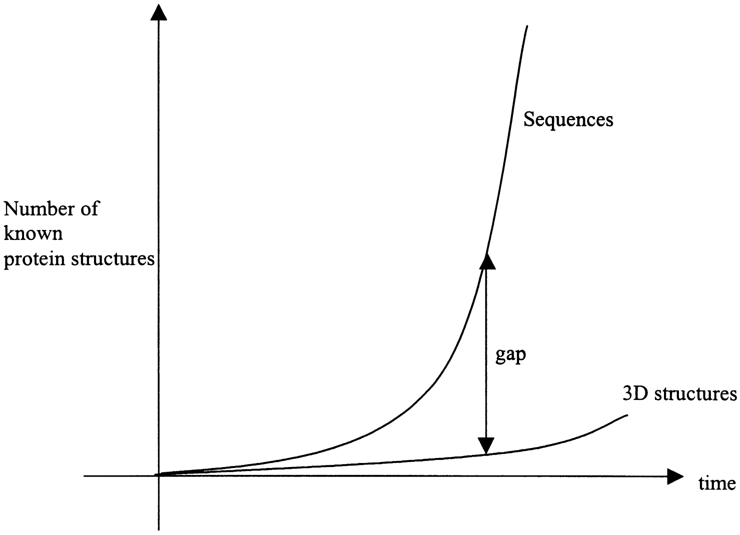


Fig. 13. Number of known sequences and number of known 3D-structures as a function of time.

started in the holistic framework. Soon after Ramachandran's publication [93,94] in 1963, force field methods arrived on the scene. Nemethy and Scheraga [106] in 1965 as well as Liquori [107] in 1969 were among the first to utilize force field methods to study peptide conformations. This is closely followed by various well-known force fields [108,109]. The most significant of such methods are CHARMm and AMBER. CHARMm, or Chemistry at Harvard Molecular Mechanics, was developed by Karplus et al. [110]. It is a program for macromolecular simulations, including energy minimization, molecular dynamics and Monte Carlo simulations. It involves analytic potential energy functions, based on experimental and ab initio data. AMBER, or Assisted Model Building with Energy Refinement, was developed by Kollman et al. [111,112]. Karplus and Kollman are in fact the pioneers of this field.

It is hard to overstate the importance of these and other force field methods. Without such calculations, neither X-ray nor NMR would be at the point where they are now. In addition, these force field methods helped the statistical prediction methods of 3D or folded structures from primary sequences. Also, using molecular dynamics the time-dependent mechanism of the folding process has been studied [113,114] with the aid of force field packages. The clear advantage of this approach is the feasibility of simulating large systems with reasonable realistic representations of the solvent. Brooks et al. [115] have also carried out numerous molecular dynamics simulations and they have looked at the effect of secondary structures in the process of protein folding.

Although the progress in protein folding due to force field-based simulations is remarkable, it has become clear [116] that in comparison to gas-phase ab initio results, some of the computed relative stabilities based on force fields disagree with ab initio results, even though the geometries are comparable. In many cases different methods yield different sets of stable structures [117–119].

A recent study of a β-heptapeptide illustrates that not all the force fields can correctly predict special conformational energetics [120]. A careful assessment of the force field in folding simulations is required. In this particular example, temperature-dependent NMR and CD spectra of methanol solution experiments reported a stable secondary structure in the range of 298–393 K. On the other hand, initial force field simulations predicted a melting point temperature of about 340 K. This dichotomy was attributed to a particular parameterization of the force field. Hence, the reductionistic approach plays an important role prior to committing to many hours of CPU on costly simulations.

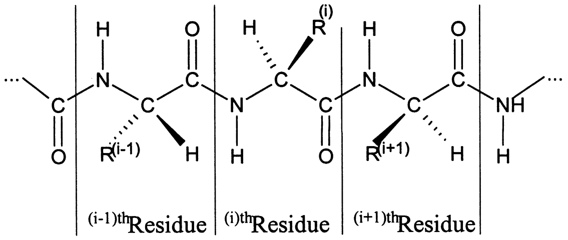
6. Internal or reductionistic modelling of protein folding

During the past 50 years, protein chemists have simplified their approach to studying protein folding by separating, at least conceptually, the problem of local backbone conformations of a single amino acid residue from that of interactions with nearest-neighbours and long-range interactions.

This line of attack implies that we must first understand the problem of backbone conformation in the absence of stabilizing, or destabilizing interactions of the side chain before we can gain a full comprehension of the entire problem. Of course, in reality, local backbone conformation includes local side chain/backbone interactions. To minimize such effects, glycine and alanine were usually used in modelling studies.

The work of Ramachandran and his group [93,94] can be regarded as the first internal or reductionistic modelling of protein folding. On the basis of these early studies, it was recognized that the flexibility of the backbone of peptide chains originates mainly from torsions about the N–C(α) and C(α)–C′ bonds (denoted by ϕ and ψ, respectively).

Studies at different levels of theory were performed to evaluate the conformational behaviour of amino acids. Extended Huckel [118,119,121, 122], CNDO/2 [121,122], PRDD [123] and PCILO [124–126] calculations were carried out in the early days. During the period of 1979–1982, ab initio calculations of peptides employed rigid (i.e. unoptimized) geometries for single amino acid diamides, dipeptides [128–132] and oligopeptides [133–136].

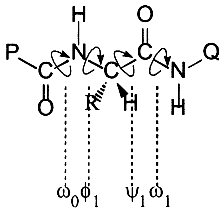


Scheme 5. A general tripepide structure in a polypeptide chain.

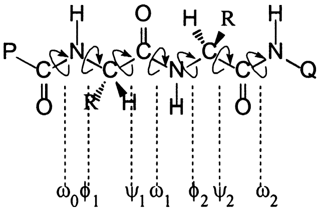
It was not until 1982 that full ab initio gradient geometry optimizations of peptides (*N*-acetyl-*N*-methylamides of glycine and alanine) were performed by Schafer and coworkers [137–139]. From these studies, it became apparent that conformational analyses with rigid geometries were not sufficient. Furthermore, conformational geometry maps [140,141] were required to give a complete description of the conformational intricacies of such systems. Subsequently, a series of similar studies [142–148] followed. Some of these included calculations with larger basis sets and single point MP2 energy calculations [147,148]. For example, MP2 gradient optimization [149] has demonstrated that unoptimized MP2 energies are potentially inaccurate [147,148]. Detailed comparisons of HF and MP2 calculations using over 10 different basis sets have also been published [150].

6.1. Peptide conformational background

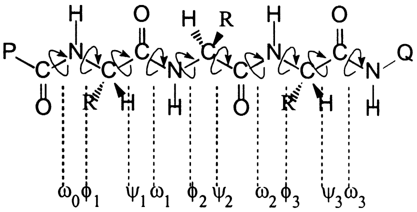
A polypeptide chain on its own, or as part of a protein molecule, consists of a series of amino acid residues as shown below in Scheme 5. Where $R^{(i)}$ specifies the side chain of the *i*th amino acid residue. Note that R may represent the side chain of the 20



I



II



III

naturally occurring amino acid residues (Table 2). Monopeptides (**I**), dipeptides (**II**) and tripeptides (**III**) contain one, two and three amino acid residues, respectively. In such models, the chain may be terminated by methyl groups or by hydrogens, as shown below for the above three cases ($P = Q = CH_3$ or H).

The torsional angles (ω_0, ϕ_1, ψ_1, ω_1,...) are responsible for folding. The energy of folding is a multi-variable function where the torsional angles are independent variables. The PEHS for the mono(1°)-, di(2°)- and tri(3°)-peptides, given below, are functions containing 4, 7 and 10 independent variables, respectively.

$$E(1°) = E[\omega_0, \omega_1, \phi_1, \psi_1] \tag{4}$$

$$E(2°) = E[\omega_0, \omega_1, \omega_2, \phi_1, \psi_1, \phi_2, \psi_2] \tag{5}$$

$$E(3°) = E[\omega_0, \omega_1, \omega_2, \omega_3, \phi_1, \psi_1, \phi_2, \psi_2, \phi_3, \psi_3] \tag{6}$$

Clearly, for a degree of polymerization of n amino acids, there are n torsional angle pairs of ϕ_i and ψ_i ($1 \leq i \leq n$), two terminal peptide functionalities (ω_0 and ω_n) and ($n - 1$) mid-chain peptide bonds [ω_i for $1 \leq i \leq (n - 1)$]. Thus, the total number of folding variables for a polypeptide containing n amino acids is

$$N = [(n-1) + 2]\omega + n\phi + n\psi = 3n + 1 \tag{7}$$

In general, the *trans* peptide bond is more stable than the *cis* peptide bond. Consequently, it has been traditional to set $\omega_i = 180°$ for all i. This limitation reduces the dimensionality of the problem substantially:

$$E(1°) = E_{trans}[\phi_1, \psi_1] \tag{8}$$

$$E(2°) = E_{trans}[\phi_1, \psi_1, \phi_2, \psi_2] \tag{9}$$

$$E(3°) = E_{trans}[\phi_1, \psi_1, \phi_2, \psi_2, \phi_3, \psi_3] \tag{10}$$

The total number of folding variables, after the reduction of the dimensionality, becomes

$$N = n\phi + n\psi = 2n \tag{11}$$

This reduction in dimensionality does not represent a negligence the importance of the *cis*-configuration of any given peptide bond, it only means that we are partitioning the problem. First, we study the backbone conformations for *trans*-peptide bonds as this represents the primary problem, and subsequently, we may study the same problem for any peptide bond in the *cis*-conformation.

Most of the study carried out so far has been centred on the generation and analysis of the $E(1°)$ potential energy surface (PES). The contour diagram of this type of PES is frequently referred to as the "Ramachandran Map", in honour of the Indian chemist, Prof Ramachandran, who first called attention to the importance of such a PES.

A topological representation of the Ramachandran map for $R = CH_3$ (i.e. for the alanine residue) is shown in Fig. 14. The various minima of the PES, representing the stable conformers, are marked by subscripted Greek letters. Although not all the 20 amino acids were subjected to ab initio computational conformational analysis, several amino acids were investigated.

The relative energies of *N*-formyl alaninamide with *trans*-peptide bond are given in Fig. 14 in the form of a PES topology. The corresponding conformational energy hyper-surface (PEHS) topology, involving both the *trans*- as well as the *cis*-isomers is depicted in Fig. 15.

As yet, these residues have been considered in the absence of conformationally variable side-chains. However, side-chains make contributions to the total energy and they are also involved in backbone/side-chain as well as side-chain/side-chain interactions, thus they ultimately help to determine protein folding. To include these further degrees of freedom in the analytic evaluations requires the extension of the above equations. Labelling of the torsional angle variables in the side-chain is accomplished using χ_1, χ_2, χ_3,... and so on, beginning at the C_α. As the side-chain is the only differentiating structural element between amino acids, each analytic equation also becomes unique. The equations then become:

$$E(1°) = E_{trans}[\phi_1, \psi_1, (\chi_1^1, \chi_2^1, ..., \chi_k^1)] \tag{12}$$

$$E(2°) = E_{trans}[\phi_1, \psi_1, (\chi_1^1, \chi_2^1, ..., \chi_k^1), \phi_2, \psi_2, (\chi_1^2, \chi_2^2, ..., \chi_k^2)] \tag{13}$$

$$E(3°) = E_{trans}[\phi_1, \psi_1, (\chi_1^1, \chi_2^1, ..., \chi_k^1), \phi_2, \psi_2, (\chi_1^2, \chi_2^2, ..., \chi_k^2), \phi_3, \psi_3, (\chi_1^3, \chi_2^3, ..., \chi_k^3)] \tag{14}$$

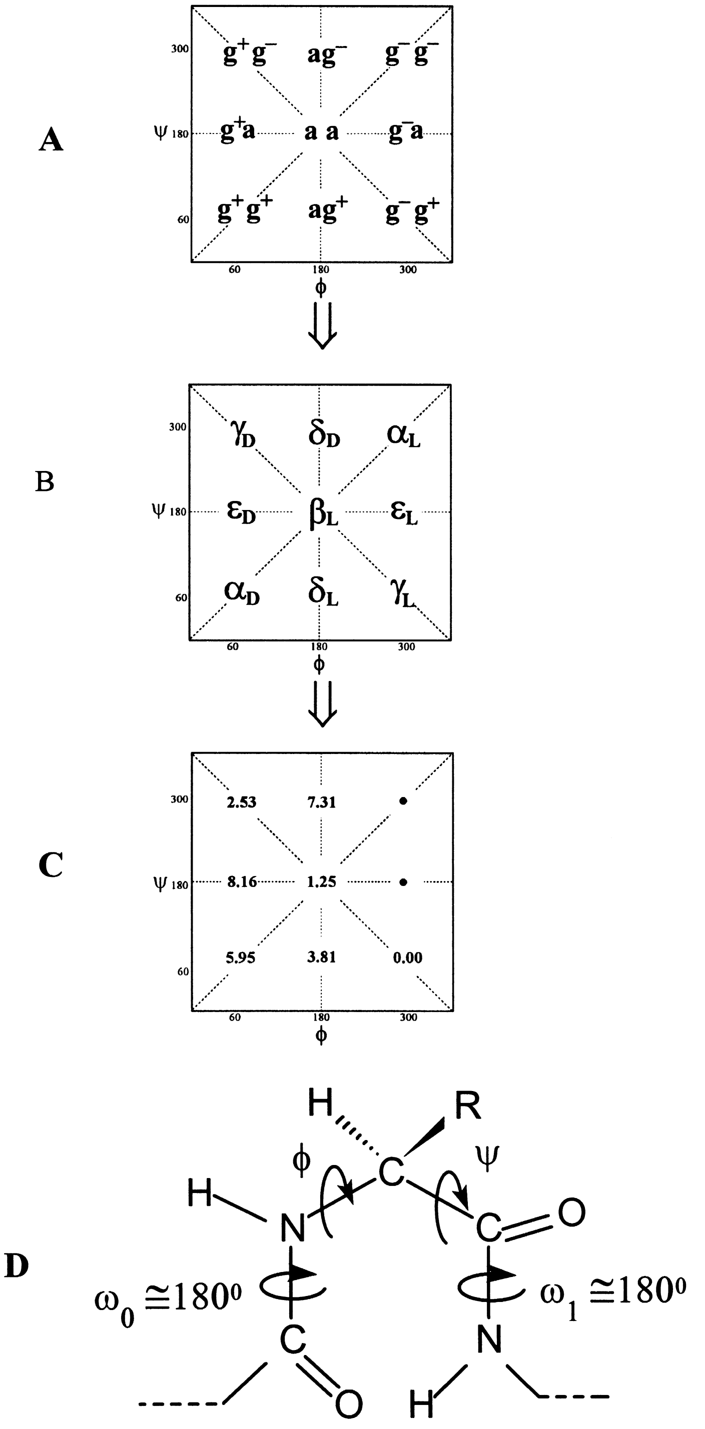


Fig. 14. Topology of a 2D-Ramachandran map: (A) conformational assignments; (B) names of conformers; (C) relative energies (kcal/mol) of *N*-formyl-L-alanineamide computed at HF/3-21G level of theory; and (D) structure of a general amino acid residue with the relevant torsional angles.

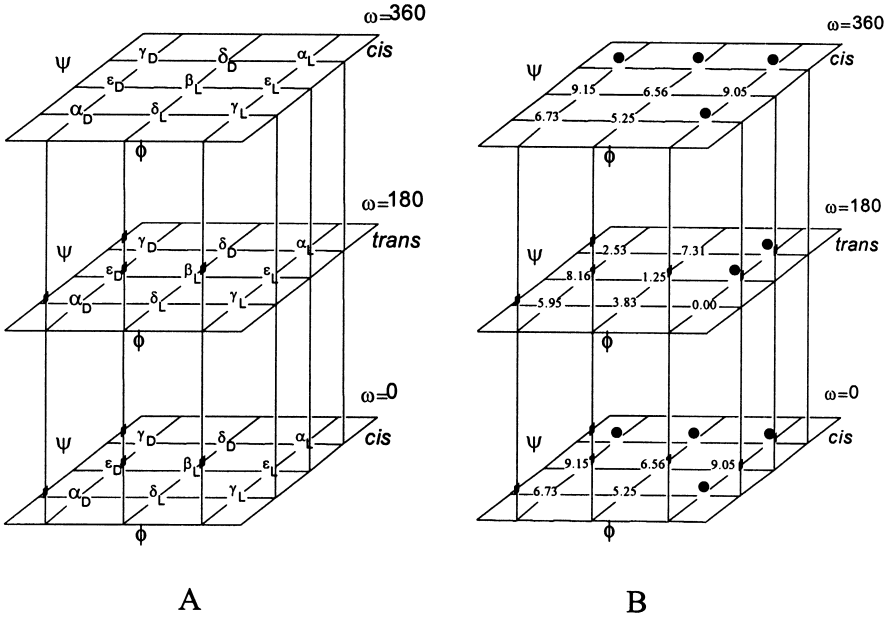

ω=360
cis
ψ
γ_D δ_D α_L
ε_D β_L ε_L
α_D δ_L γ_L
ϕ
ω=180
trans
ω=0
A
9.15 6.56 9.05
6.73 5.25
2.53 7.31
8.16 1.25
5.95 3.83 0.00
B

Fig. 15. Topology of a 3D-Ramachandran PEHS: (A) name of conformers; (B) relative energies of HCO–(ω)–NH–(ϕ)–CHMe–(ψ)–CONH$_2$.

where (χ_1, χ_2,…,χ_k) is specific to each amino acid and its side-chain, with the superscript denoting the residue that side-chain belongs to in the polypeptide chain.

Putting this concept to a generalized form, we obtain the following multi-variable function:

$$E(n^\circ) = E_{trans}[\phi_1, \psi_1, (\chi_1^1, \chi_2^1, ..., \chi_k^1), ..., \phi_n, \psi_n, (\chi_1^n, \chi_2^n, ..., \chi_k^n)] \quad (15)$$

6.2. Single amino acid diamides

There are two ways to represent the rotation about a single bond. Traditionally, the 0 → 180 → 360° range is used, but recently, IUPAC recommended the convention of −180 → 0 → +180°This convention has the advantage of designating the 0 → +180° segment as clockwise rotation and the 0 → −180° segment as a counter-clockwise. However, it has the disadvantage of certain minima falling on the edges or the corners of the Ramachandran map. The two representations are presented in Fig. 16.

Peptide models, such as CH$_3$CONH–CHR–CONHCH$_3$ or simply HCONH–CHR–CONH$_2$ can mimic the *i*th amino acid residue in a protein chain. The ϕ and torsional angles are defined in Fig. 14. The conformational assignments (g$^+$g$^+$, g$^+$a,…, etc.) are shown in Fig. 14.

The names of the minima (Fig. 14) are subscripted Greek letters. The Greek letters originate from earlier nomenclature (involving α, β and γ) while the L and D subscripts originate from the observation that L-amino acids favour L conformations while D-amino acids favour D conformations (lower part of Fig. 17). The names also suggest the combination of the chirality of a constitutional structure (*R* or *S* configuration) and that of the conformational twist or folding. This is summarized in Fig. 18.

The top of Fig. 19 shows the symbolic representation of a conformational PES for two full cycles of rotation (+360 → 0 → + 360°) of both ϕ and ψ. The PES can be partitioned into four quadrants, in the traditional way, or it can be partitioned according to IUPAC convention, as shown by the broken lines.

An energy contour diagram of the conformational PES for a peptide (PCONH–CHR–CONHQ), presenting two full cycles of rotation (−360 → 0 → 360°) of both ϕ and ψ, is shown at the bottom of Fig. 19. The central square is the

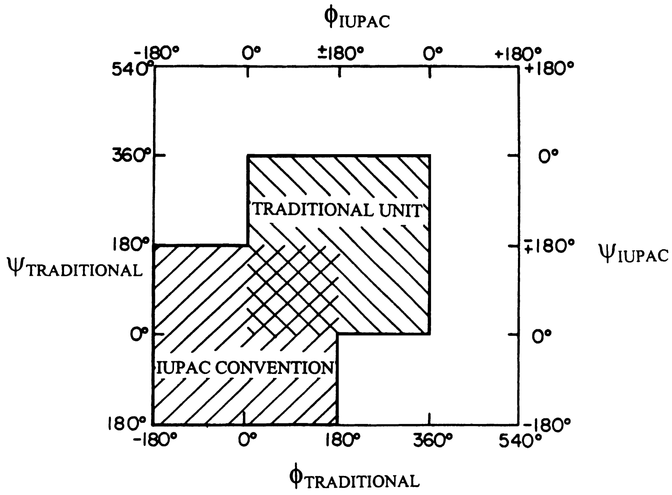


Fig. 16. Two kinds of partition of the PES. The central square corresponds to the traditional cut (0 → 180 → 360°), while the lower left-hand square represents the IUPAC conventional cut (180 → 0 → + 180°).

IUPAC conventional cut, while the four quadrants are the traditional cuts. One of these traditional cuts (e.g. the upper right-hand quadrant) is shown in pseudo-3D-representation in Fig. 20.

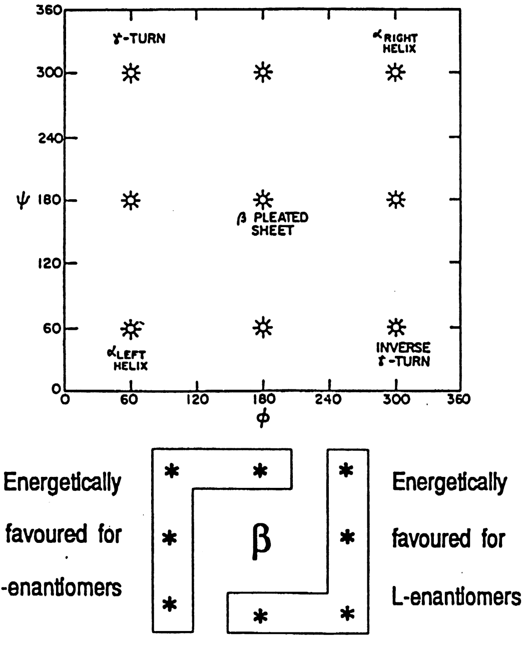


Fig. 17. Underlying principles for choosing subscripted Greek letter (e.g. α_L, α_D, β_L..., etc.) as names for the peptide conformations.

Eighteen out of the 20 naturally occurring amino acids have the same type of backbone folding as shown in Fig. 14 (i.e. nine discrete conformations). The two exceptions are proline and glycine.

Proline's nitrogen is locked in a five-membered ring. For proline residue (**IV**), ϕ can only be in the vicinity of $-60°$ (i.e. $+300°$) and, therefore, only three backbone conformations are possible: α_L, ϵ_L

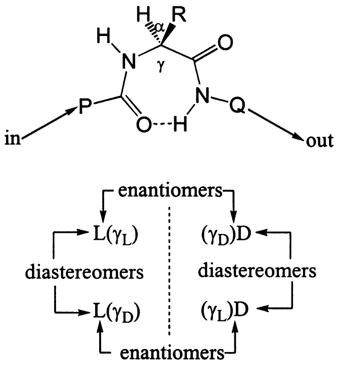


Fig. 18. Stereochemical relationships of γ-turns. Note, that not only the α-carbon has chirality, but there is also chirality in the twisting of the backbone. The combination of these two types of chiralities leads to enantiomeric and diastereoisomeric structures. D and L denote the chirality of the C^{α} configuration, while γ_L and γ_D denote the chirality of the conformation.

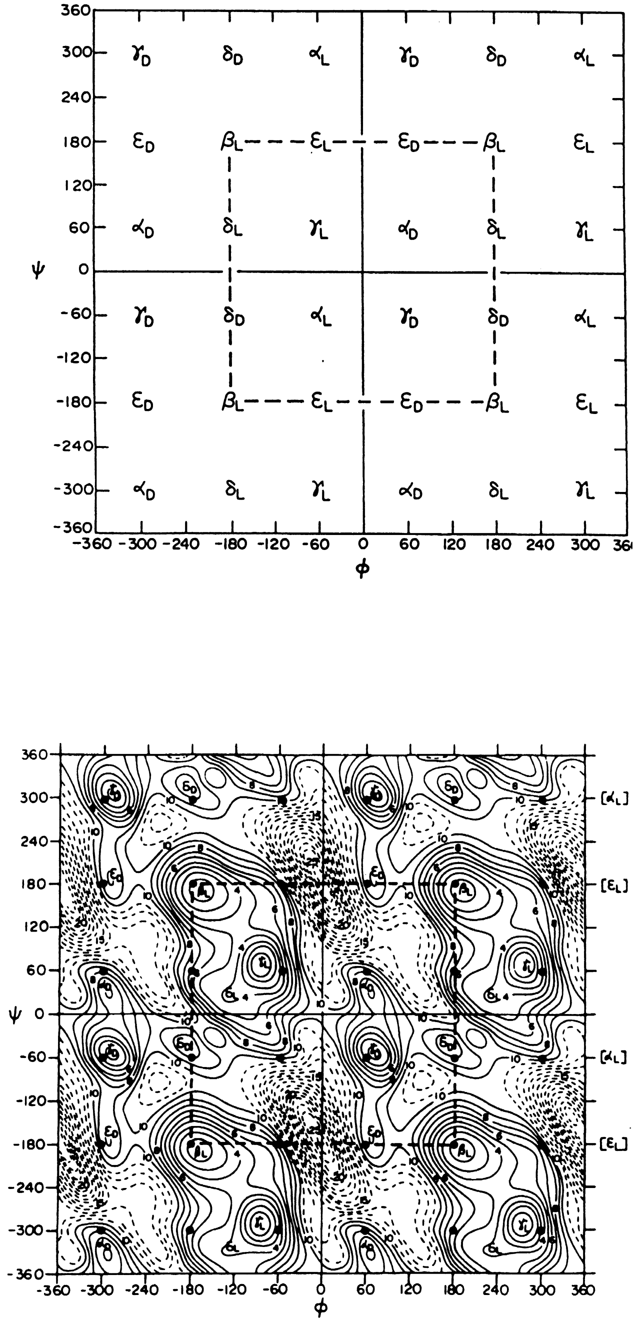


Fig. 19. Top: a schematic representation of the conformational PES of a peptide (PCONH–CHR–CONHQ). Subscripted Greek letters symbolize the approximate locations of the conformations. Bottom: contour diagram of the 2D-Ramachandran potential energy surface of HCONH–CHCH$_3$–CONH$_2$, presented in the $-360 \leq \phi \leq 360°$ and $-360 \leq \psi \leq 360°$ range of independent variables. The central square (broken lines) is the IUPAC conventional cut, while the four quadrants are the traditional cuts.

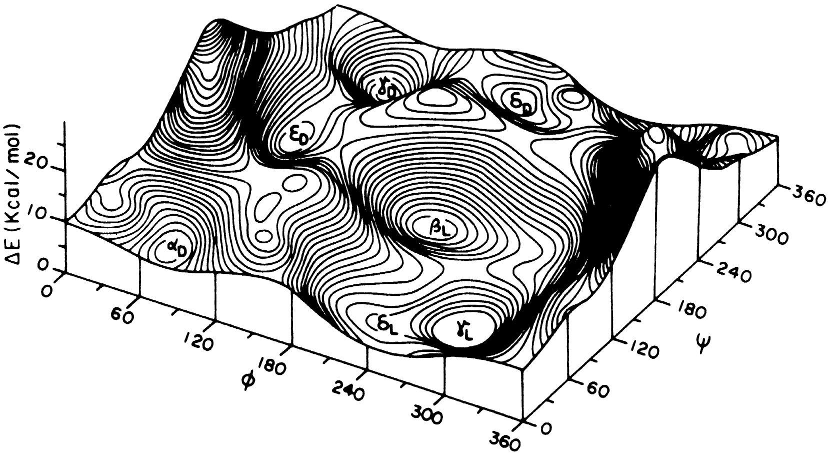
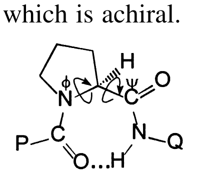
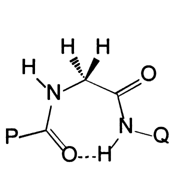


Fig. 20. Pseudo-3D-Ramachandran PES of HCONH–$CHCH_3$–$CONH_2$, presented in the $0 \leq \phi \leq 360°$ and $0 \leq \psi \leq 360°$ range of independent variables. This represents one of the four equivalent quadrants in Fig. 7.

and γ_L. The other unique amino acid is glycine (**V**), which is achiral.



IV



V

Proline is fundamentally different from all the other 18 chiral amino acids in more than one respect:

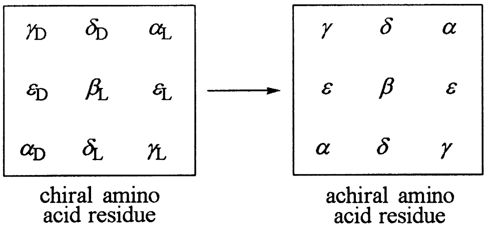
1. the R group forms a five-member ring with the backbone;
2. there is no peptidic N–H group in the residue to be involved in hydrogen bonding;
3. since there are two carbon atoms connected to the nitrogen, there is a greater chance of *cis/trans* isomerization in the peptide bond.

The potential energy cross-sections of the type $E = E(\psi)$, for the Ramachandran map of HCO–Pro–NH_2 containing *cis*- and *trans*-peptide bonds [162] are shown in Fig. 23. Preliminary investigation on the *cis*-peptide bond has been completed and the *cis*-Ramachandran map is currently under construction [3].

In the case of the glycine residue, double degeneracy occurs in its conformational PES as shown in Fig. 21. Finally, it should be mentioned that for certain molecular residues, molecular computations have determined the actual locations of the nine minima shown above. There are small deviations from the ideal ϕ and ψ values. Table 4 lists these numerical values for alanine residue. The information tabulated above is also presented graphically in Fig. 22.

Ab initio SCF calculations as well as DFT calculations allow us to compute the energy for the molecule as a whole as well as selected fragments of that molecule. In doing so, one can evaluate partial contributions to the total energy, and consequently analyse rigorously the different factors involved, for example, in backbone/side-chain interactions. In this way, it is possible to calculate, using *isodesmic* reactions, the stabilization energy exerted by the side chain on the backbone of the amino acid residue.

On the basis of the aforementioned analyses, one can predict, at least in a semi-quantitative way, the effects exerted by a given side chain. It is also possible



γ_D	δ_D	α_L
ε_D	β_L	ε_L
α_D	δ_L	γ_L

chiral amino acid residue

γ	δ	α
ε	β	ε
α	δ	γ

achiral amino acid residue

Fig. 21. The development of double degeneracy of the PES when a chiral amino acid residue is changed to an achiral amino acid residue.

Table 4
Optimized ϕ, ψ torsional angle pairs for alanine residue. The idealized torsional angle pairs, together with their conformational classification, are also shown for the sake of comparison

Conformer	Optimized values		Idealized values		Conformational classification
	ϕ	ψ	ϕ	ψ	
α_L	−66.6	−17.5	−60	−60	g^-g^-
α_D	+61.8	+31.9	+60	+60	g^+g^+
β_L	−167.6	+169.9	−180	+180	aa
γ_L	−84.5	+68.7	−60	+60	g^-g^+
γ_D	+74.3	−59.5	+60	−60	g^+g^-
δ_L	−126.2	+26.5	−180	+60	ag^+
δ_D	−179.6	−43.7	−60	−60	ag^-
ϵ_L	−74.7	+167.8	+60	+180	g^-a
ϵ_D	+64.7	−178.6	−180	−180	g^-a

to study the influence of a portion of side chain to other parts of the peptide or protein as they come in each other's proximity. This method can be applied to any kind of natural or modified amino acids and is expected to contribute to a better understanding of some less noticeable effects, which might strongly influence the structure of a polypeptide or a protein.

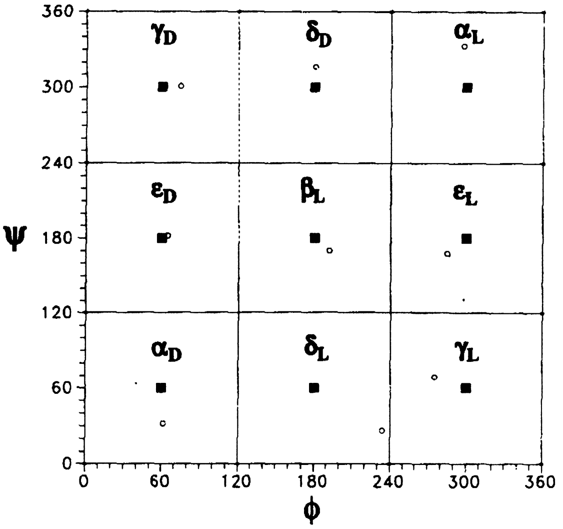


Fig. 22. A schematic illustration of the PES of an average amino acid residue, obtained from the calculations carried out so far on mono-, di- and tri-peptides. The idealized positions are marked by shaded squares and the computationally determined positions are shown as open circles. The names of the conformers are given as subscripted Greek letters. Note that a single amino acid residue might not be able to take on all of the shown conformations.

The topological representation of the conformational PES (Ramachandran map) of For–Ala–NH_2 is shown in Fig. 14, while that of the conformational PES is shown in Figs. 19 and 20. Work has been completed for the following N- and C-protected amino acids containing a *trans*-peptide bond: Gly [142–148], Ala [142–148], Val [151], Phe [152,153] and Ser [154–156]. Preliminary studies have been published on Pro [157], Asp [158], Asn [159], Cys [160] and Sec (selenocysteine) [161]. The following protected amino acid residues, again with *trans*-peptide bonds, are currently under investigation: Arg, Lys, His, Tyr, Leu, Ile, Thr, Trp, Glu, Gln and Met.

6.3. Peptide models

Peptide models, or the "dipeptide approximation" [93,94,117–167] as it is often referred to in the literature has often been applied in the developing of parameters for empirical energy calculations on single amino acid diamides.

In this method, it is assumed that the values of ϕ_i and ψ_i in the ith residue of a peptide chain depend mainly on one another and on the nature of the residue R_i. However, the ϕ, ψ values are expected to be largely independent of the neighbouring pairs of ϕ_{i+1}, ψ_{i+1}, ϕ_{i-1} and ψ_{i-1}. The model implies that essential conformational properties of polypeptides may be deduced from their isolated components.

This approach has been quite successful in describing peptide conformational properties since short-range interactions are dominant in the folding of a

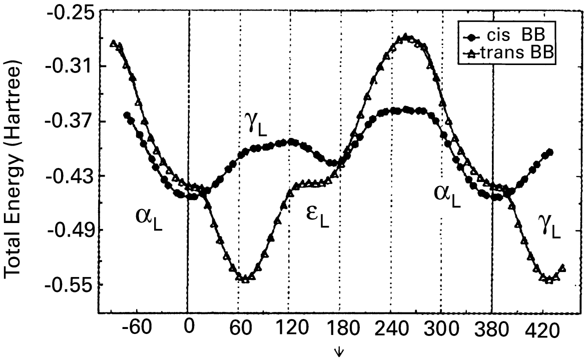


Fig. 23. Conformational potential energy curves $E = E(\psi)$ for *cis*- and *trans-N* formyl-L-prolinamide.

polypeptide chain. At the same time, the model neglects cooperative phenomena in polymers and long-range interactions between groups that are remote from each other along the backbone of the peptide chain. Thus, characteristic differences in the properties of dipeptides and polypeptides were also found. For example, in contrast to many empirical potential energy studies, it is apparent from ab initio geometry optimizations of model dipeptides that the right-handed-helical conformations, are not minima in the single amino acid diamides, even though they are common in proteins [116,146,168].

In pursuing the main goal of peptide conformational analysis, i.e. computing a 3D structure of a polypeptide from its amino acid sequence, extensions beyond the small-peptide model is required. Thus, it is interesting to see how the properties of oligopeptides deviate from the sum of the properties of the component single residues. For some conformations, among them turns, bends and helices, stabilization may arise from interactions between different single residues. Hence, model calculations beyond peptide approximation are needed for investigating such effects. In the past,

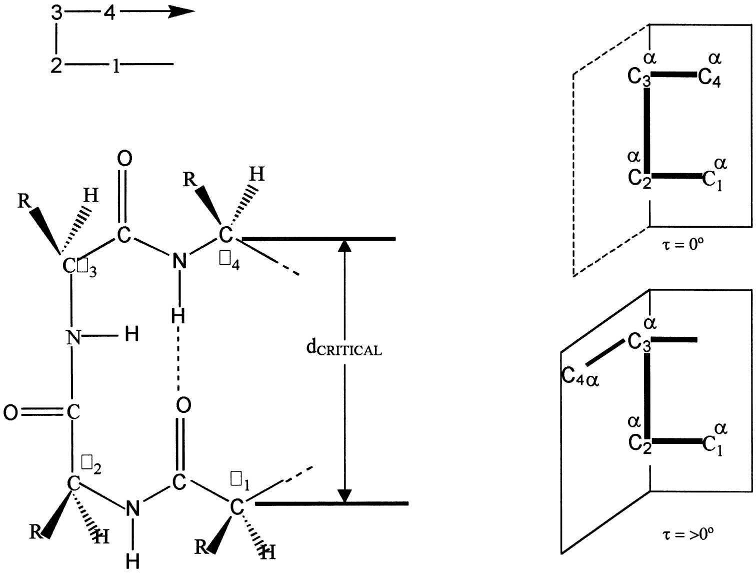
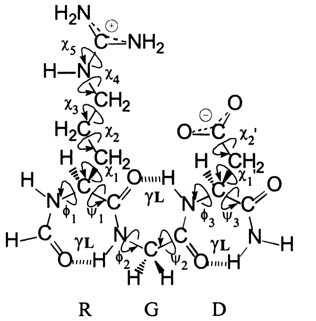
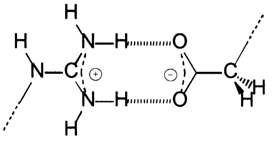


Fig. 24. A schematic illustration of hairpin or β turn conformation. Left: untwisted molecular structure; right top: untwisted molecular structure; right bottom: twisted molecular structure. The extent of twist is denoted by τ.

VI



VII



such calculations were mainly performed with empirical procedures [169], but there were also early attempts to identify cooperative effects in oligopeptides by using ab initio calculations [136]. By necessity, the latter were rigid geometry calculations. This was a disadvantage, recognized by their authors, because such analysis can only be partially successful without geometry optimizations.

6.4. Dipeptide diamides

N- and C-protected dipeptides represent an important model for β-turns. Their role in defining and analysing β-turns is shown in Fig. 24. After some preliminary work [170], an analysis on HCO–Ala-Ala–NH_2 was carried out at the ab initio level [171] in 1993. A more detailed study was made available [172] in 1994. A full account of all existing minima has been published by Schafer et al. [173] in 1998.

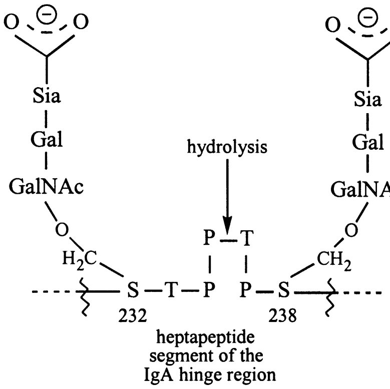
6.5. Oligopeptide diamides

Mostly alanine has been studied in tripeptide, HCO–Ala_3–NH_2 and tetrapeptide HCO–Ala_4–NH_2 forms [174–178]. Schafer and coworkers studied [179] N- and C-protected pentaglycine and pentaalanine in 1963.

Based on restricted Hartree–Fock (RHF) 3-21G calculations on peptide models of systematically increased residue length, (HCO–L-(Ala)$_n$–NH_2), the selection of some homo-conformers, such as helices and sheets, was observed. With the increasing n, the two helical forms, $(\alpha_L)_n$ and $(\alpha_D)_n$, become more strongly favoured [180]. Larger α-helices have also been studied [9] and some representative structures are shown in Fig. 25.

There are many interesting possibilities in tri- and tetrapeptides once we include amino acids beyond glycine and alanine. For examples RGD (Arg-Gly-Asp) and PPTP (Pro-Pro-Thr-Pro) have been studied in a preliminary fashion [3] and have already revealed many unexpected features. RGD (**VI**) shows an internal salt bridge (**VII**) in the absence of external neutralization.

When the carboxylate moiety is coordinating with Ca^{2+} ion the chain becomes extended. Typical optimized structures [3] are shown in Fig. 26. PPTP is a segment of the hinge region of immunoglobulin A (**VIII**) which may undergo proteolytic cleavage during the immune process. An optimized structure [3] for PPTP is shown in Fig. 27.



VIII

Clearly a great deal can be learned from oligopeptide studies concerning backbone/side chain, and side chain/side chain interactions. Such studies will undoubtedly be the basis for the reductionistic approach.

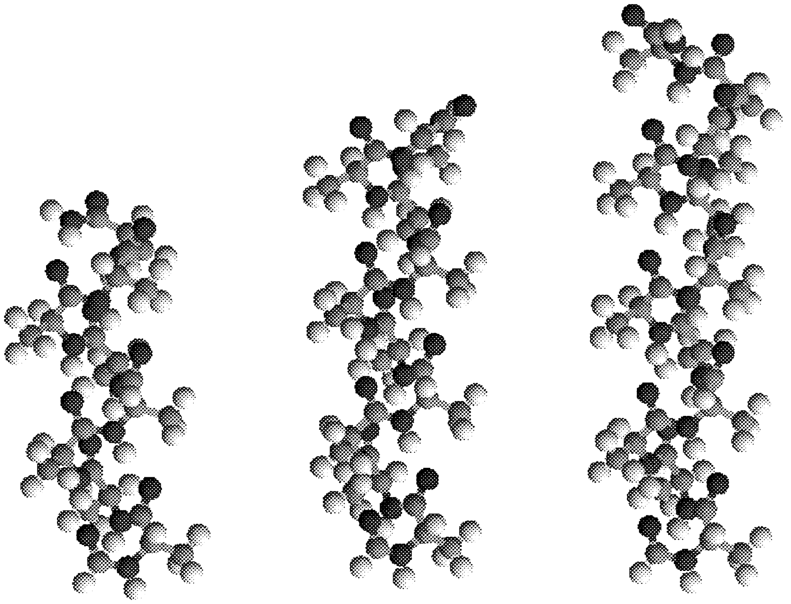


Fig. 25. Helical structures of HCO–(Ala)$_4$–NH$_2$ (for $n = 8$, 10, 12) optimized at HF/3-21G level of theory.

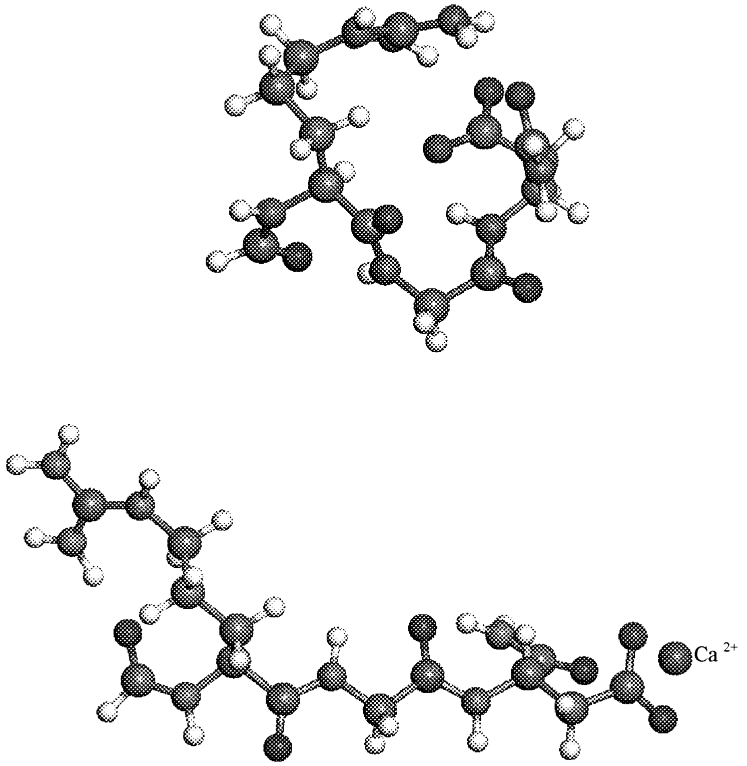


Fig. 26. Two conformers of RGD without (top) and with Ca^{2+} (bottom) optimized at the HF/3-21G level of theory.

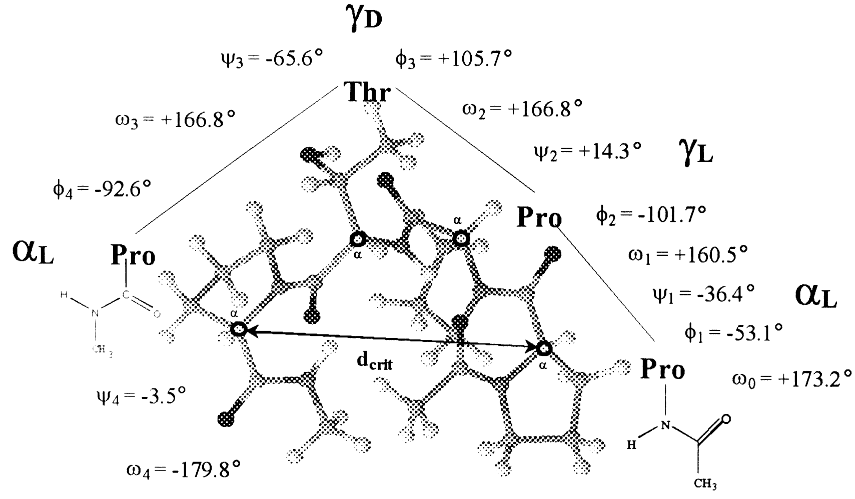


Fig. 27. An optimized structure of the tetrapeptide Pro-Pro-Thr-Pro (PPTP).

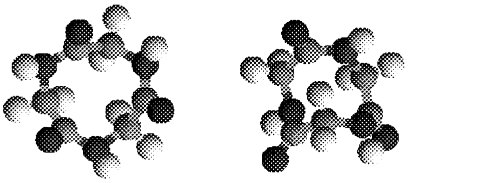


Fig. 28. Ab initio structures of all-*cis* cyclic triglycine optimized at HF/3-21G level of theory.

6.6. Cyclopeptides

Cyclopeptides cannot be hydrolysed by proteolytic enzymes. This property makes them good drug candidates as they can be designed to be inhibitors of enzymes, or to block the active sites of receptors. In contrast to linear peptides where all the conformations may be predicted by the use of the rules of multidimensional conformational analysis (MDCA), cyclopeptide conformations remain a mystery.

It is noteworthy that besides linear peptide models, cyclic forms, such that dioxopiperazine, cyclotetra-, cyclopenta- and cyclohexa- peptides, have also been investigated by means of ab initio techniques. For examples, 3-21G and DZP RHF calculations [181] have been used for the determination of low-energy conformers of cyclohexaglycine.

Our own optimized structures are shown in Figs. 28–30.

7. Future prospects

It is now useful to summarize where we are at in our reductionistic approach concerning peptide and protein folding. We have witnessed the publication of ab initio conformations of about half of the single amino acid diamides. The computations on the other half is now in progress. This portion of the conformational analysis of peptides, that is, the study of the 20

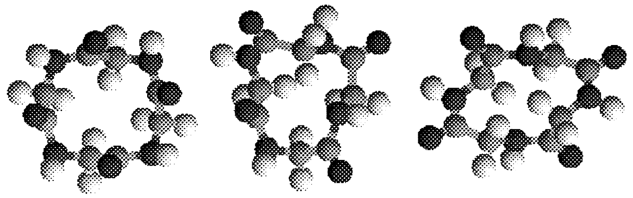


Fig. 29. Ab initio structures of all-*cis* cyclic tetraglycine optimized at HF/3-21G level of theory.

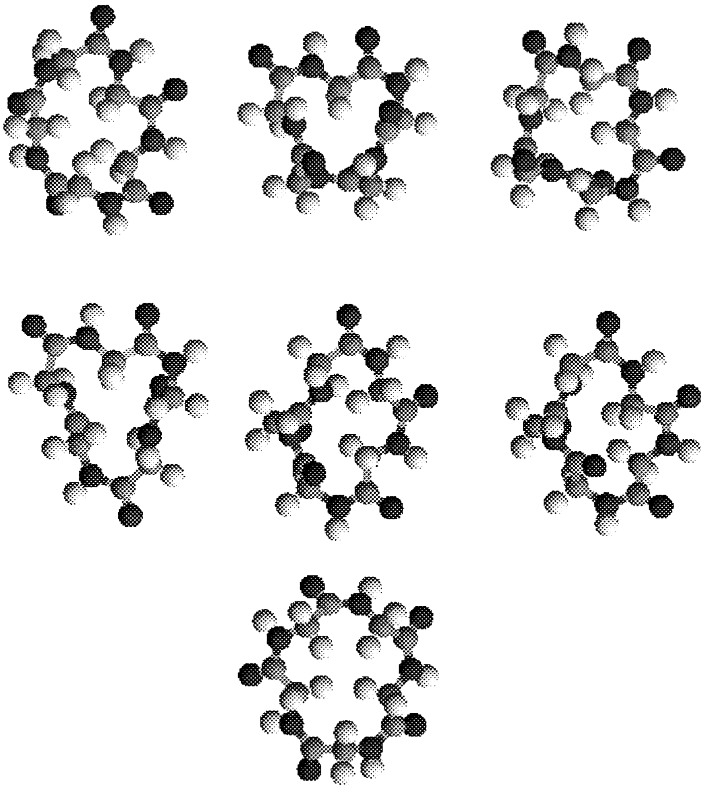


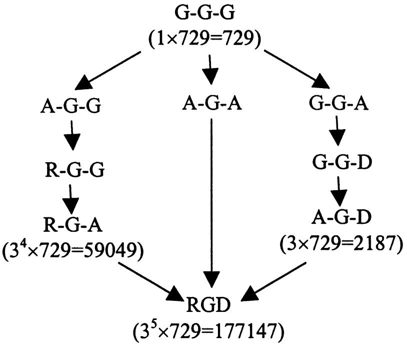
Fig. 30. Ab initio structures of all-*cis* cyclic pentaglycine optimized at HF/3-21G level of theory.

N- and C-protected amino acids, will soon be completed.

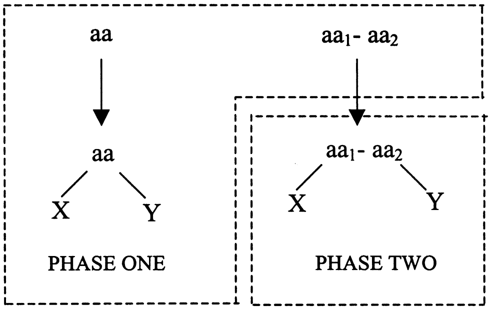
The next phase includes dipeptides (diamino acid diamides) and tripeptides (triamino acid diamides). From the 20 amino acids, we may generate $20^2 = 400$ primary sequences of dipeptides and $20^3 = 8000$ primary sequences of tripeptides. Of course, each of these structures has many backbone and side chain conformations. The Ramachandran PES associated with each amino acid residue has up to nine minima as discussed earlier. This gives rise to $9^2 = 81$ and $9^3 = 729$ backbone conformations for dipeptides and tripeptides, respectively. The side chains may have various conformations. Glycine has no side chain at all, and the methyl side chain of alanine has only one conformation. Thus, for practical purposes, systems such as the ones shown below in Eq. (16), have only backbone conformations as there are no side chain variations.

Gly-Gly-Gly   Ala-Ala-Ala   (16)

However, as discussed earlier in a peptide like Arg-Gly-Asp (or RGD), the backbone (**VI**) may be represented by a 6D-conformation subspace ($\phi_1, \psi_1, \phi_2, \psi_2, \phi_3, \psi_3$). This leads to $3^6 = 9^3 = 729$ backbone conformations. The full representation of the side chain is a 7D-conformational subspace ($\chi_1, \chi_2, \chi_3, \chi_4, \chi_5, \chi_1', \chi_2'$). Since we may expect only one distinct conformer along χ_5 and χ_2', the side chain may be represented by a 5D-conformational chain subspace ($\chi_1, \chi_2, \chi_3, \chi_4, \chi_1'$). Consequently, we may expect $3^5 = 243$ side chain conformations. The total number of distinct RGD conformers may, therefore, be $729 \times 243 = 177\,147$. The RGD molecule is interesting on its own right because the positively and negatively charged side chains may be



Scheme 6. A sequential increase in complexity on going from the simplest GGG to the more complicated RGD motif.



Scheme 7. A sequential increase in complexity on going from monopeptide aa to tripeptide Y-aa-Y, and from dipeptide aa1-aa2 to tetrapeptide X-aa1-aa2-Y.

engaged in salt bridge formation (**VII**) either via an intra- or an inter-molecular connection.

However, from our present point of view, RGD can demonstrate possible side chain/side chain interactions. In addition, one can show how the central neutral glycine is affected by its nearest-neighbours, arginine and aspartic acid. Such nearest-neighbour interactions can be assessed by comparing the structures in Scheme 6 (number of conformers are shown in parentheses).

Some tripeptides have many more stable conformations than the 177 147 structures associated with RGD. For example, RGR is expected to have up to $3^8 \times 729 = 6561 \times 729 = 4\,782\,969$ conformers. Clearly, we are dealing with a rather large number of structure determinations. Even if we assume, on average, only 10^5 conformers for a tripeptide, and 8×10^3 primary sequences, then we must have roughly $8 \times 10^3 \times 10^5 = 8 \times 10^8$ (eight hundred million) geometry optimizations.

If we could have at our disposal 1 million processors, then each of these processors must carry out 800 optimizations.

If a single optimization could be carried out, on average, in 1 min, then 800 optimizations could be performed in half a day.
If the average processing time were to be 1 h then the whole process might take one month.
If, however, the average processing time is 1 day, then, the 800 geometry optimizations will require slightly more than 2 years.

Thus, the time requirement seems to be manageable, if only we could have 1 million very fast processors (Fig. 31).

We could consider the mono-, di- and tri- peptides analyses as PHASE ONE of the reductionisitc approach (Scheme 7). The above numbers may look frightening in the year 2001, but within a few years computer hardware and software development will make such a computational project a relatively minor undertaking. With the completion of PHASE ONE, however, we could understand fairly well the nearest-neighbour interaction.

PHASE TWO (Scheme 7) should cover the topic of tetrapeptides. In a tetrapeptide, we are sandwiching a dipeptide (i.e. diamino acid unit) between two nearest-neighbours as depicted below (Scheme 7) where the symbol, aa, denotes an amino acid at central positions and X and Y symbolize nearest-neighbouring amino acids.

In the case of tetrapeptides, we have $20^4 = 160\,000$ primary sequences. Each of these sequences (i.e. every possible tetrapeptide structure) could have up to $9^4 = 6561$ backbone conformations. Even if (in accordance with Table 2) each of the four side chains may be limited to double rotors (i.e. Leu, Ile, Ser, Thr, etc.), there are still $(3^2)^4 = 3^8 = 6561$ side chain conformations for a tetrapeptide. Thus, the total structural determination for all tetrapeptide structures is expected to be of the order of $(6561)^2 = 43\,046\,721$ geometry optimizations. Consequently, at the current rate of computability PHASE TWO would require several centuries. If a tetrapeptide optimization such as MeCO–$(\text{Ala})_4$–NHMe at the HF/3-21G level of theory for a single

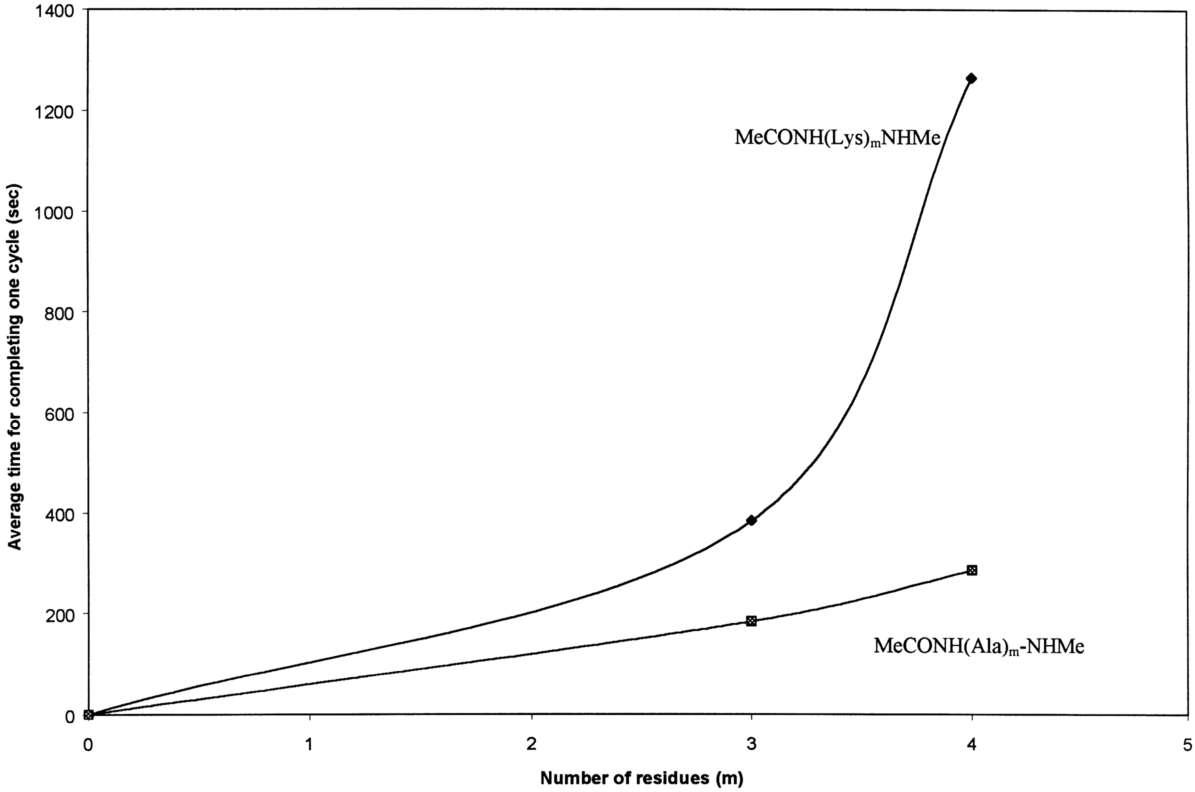


Fig. 31. Average time for computing one cycle for the $[\beta_L]_n$ conformers of MeCO–NH[Xxx]$_n$–NHMe where Xxx = Ala or Lys.

conformation could take something of the order of 1–10 days, the whole project would require 43 046 721 to 430 467 210 days. Thus, revolutionary hardware and software development would be required to complete PHASE TWO within a reasonable timeframe. The questions are, however, how soon will such hardware development come, and to what extent will it improve computability. According to Fig. 4, the answer is the year 2025, the pessimistic (i.e. the lower) curve suggests a 10^6-fold increase (from 10^{11} to 10^{17} FLOPS) and the optimistic (i.e. the higher) curve indicates a 10^{12}-fold increase (from 10^{11} to 10^{23}). Taking into consideration of only the pessimistic prediction (i.e. 10^6-fold increase), the 430 467 210 days will be reduced to 430 days, or optimistically speaking 430 days. Such computability will be more than adequate. Thus, within a quarter of a century, i.e. within the professional lifetime of our students, the data generation will be feasible.

The various time units (for example 1 min, 1 h, 1 day and 10 days) used in the previous "thought-experiment" is an arbitrary way to illustrate a point. However, it would be nice to know, as initial conditions, how fast can we compute today the smallest tri- and tetrapeptide (Ala-Ala-Ala and Ala-Ala-Ala-Ala). Also it would be nice to know how the computations increase if we use the amino acid with the longest side-chain: lysine (Lys). For this reason Lys-Lys-Lys and Lys-Lys-Lys-Lys were also computed and shown in Table 5 for the sake of comparison. Of course computations at a level of theory higher than HF/3-21G take considerably longer time than those shown in Table 5. All the calculations in Table 5 were carried out on a CRAY T3E 600. Optimization calculations using 32 processors and are summarized in this table. As previously noted for this size of systems and basis sets, the efficiency is approximately 60% [182]. Although the CRAY T3E 600 has an older EV5 microprocessor, the relative increase in time in going from Ala-Ala-Ala to Lys-Lys-Lys or Ala-Ala-Ala-Ala to Lys-Lys-Lys-Lys should be proportional

Table 5
A comparative benchmark calculation for the smallest $(Ala)_3$ and $(Ala)_4$ as well as largest $(Lys)_3$ and $(Lys)_4$ N- and C-protected tri- and tetrapeptides computed at HF/3-21G level of theory

Peptide	Conform	CRAY T3E 600[a] E (Hartree)	m[b]	t_{total} (s)	t_{total}/m (s)[c]
Ala-Ala-Ala	α	−979.0962440	39	8663	222.1
	β	−979.0936655	55	10094	183.5
Lys-Lys-Lys	α	−1493.6400223	40	27632	690.8
	β	−1493.6590670	106	72525	684.2
Ala-Ala-Ala-Ala	α	−1223.5897045	45	16518	367.1
	β	−1223.5815535	57	16286	285.7
Lys-Lys-Lys-Lys	α	−1909.5579932	53	73719	1390.9
	β	−1909.5544625	54	67610	1252.0

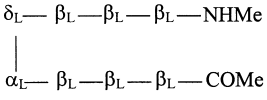
[a] 32 processors are used.
[b] m = number of optimization cycles.
[c] average time for completing one cycle.

Table 6
Number of atoms (N) in small oligopeptides MeCO–NH–$[Xxx]_n$–NHMe where residue Xxx maybe Ala or Lys. Note that the number of variables to be optimized is $3N - 6$

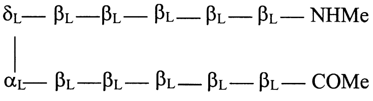
n	N		$3N - 6$	
	$[Ala]_n$	$[Lys]_n$	$[Ala]_n$	$[Lys]_n$
3	42	78	120	228
4	52	100	150	294
5	62	122	180	534

with faster microprocessors or the fastest vector machine. As we go from Ala-Ala-Ala to Lys-Lys-Lys or Ala-Ala-Ala-Ala to Lys-Lys-Lys-Lys, we not only increase the size of the system it is comprised of (see Table 6) but we have increased the number of basis functions, from 224 to 380 or 279 to 487, respectively. In terms of CPU usage, going from Ala-Ala-Ala to Lys-Lys-Lys represents more than a factor of 3. Clearly, conventional ab initio methods scale up very quickly. Future calculations will need to rely more and more on linearly scaling methods. Recently, Scuseria et al. [183] fully optimized a molecular system with 1226 atoms at the PM3 level of theory.

We could be confident that before 2025, not only can we solve the tetrapeptide structures, but we can also study secondary structural motifs, such as anti-parallel β-pleated sheets and helical structures. This could be regarded as PHASE THREE. It should be noted that PHASE ONE and PHASE TWO could be completed with a distributed mode of computation where each conformation will be processed by a separate processor, however, for PHASE THREE, parallel processing will be necessary for the increased size of the problem. Since these motifs represent unique backbone conformations, only the side chain orientations are required to be optimized extensively. We probably would need a minimum of eight amino acids in the peptide chain, but one can envisage that 12 amino acids could be treated. We have already optimized [9] a dodecaalanineamide helix: HCO–$(Ala)_{12}$–NH_2, in 1999. At this point, we noticed that the solvent effect does have a significant effect on the stabilities of the α-and the 3_{10} helices. Consequently, the inclusion of solvent effect will be important in the future. For the anti-parallel β-pleated sheet, the octa- and dodeca-peptides would look like the segments shown below:



IX



X

If we reach manageable computability for such sizes before 2025, we will have numerous optimized structures in a gigantic database. No human being is capable of analysing manually such a gigantic pile of structural data involving various geometries and stabilities. Thus, while we are building our *ant-hill-like database*, we also have to develop algorithms to study the results automatically. From such database, we can learn the extent of various side-chain/backbone and side chain/side chain interactions and their geometrical boundary conditions of occurrences. We could fit mathematical functions to their conformational hypersurfaces where energy is a function of geometry. With such analytic functions, numerical predictability will be within reach. Thus, protein folding will become a problem in the field of mathematical analysis of multivariable functions. Needless to say, the number of variables for such an analytic function is very large. At that time, we will be armed sufficiently to attack the "dragons", and will no longer have to run away from them, as Ira Ramsen suggested in the beginning of the twentieth century. We will be able to construct at that time numerically reliable model which we could only name at this time as "super force fields". When this milestone is upon us, the period of reductionistic approach will be complete, and the advent of the next holistic period will begin.

How long does it take before we reach this "Promised Land"? Are we going to arrive there by 2030, 2040 or 2050? One can only speculate. Nevertheless, we can be confident that well before the end of this century, we will know the secret of the forces that dictate the folding of a given protein. Our younger colleagues are impatient; they are eager to know if we are there yet. We are not there as yet, but hopefully, with favourable conditions, they will probably be there before their retirements. Pessimistically speaking, with unfavourable conditions, the solution will be in the hands of their students.

We can give them only a general advice:

> Be prepared to embrace the future, because the future is coming and the future usually arrives sooner than we are ready to give up the present.

Acknowledgements

The authors are grateful to Prof. A. Kucsman for helpful discussion. We wish to acknowledge the cooperation of Dr Igor A. Topol at the National Cancer Institute (NCI) in the polyalanine helix project.

One of the authors (I.G.C.) wishes to acknowledge the continuous support of the National Research Council (NRC), and its funding-successor, the National Sciences and Engineering Research Council (NSERC) of Canada for the period of 1964–1999. We would like to thank the CRAY Inc. Computer Centre in Eagan for allowing us to run the benchmark calculations on a CRAY T3E-600. We are also grateful to the National Cancer Institute (NCI) for the use of services at Frederick Biomedical Supercomputing Centre. Finally, we also wish to thank Velocet Communications Inc. for database management, network support, software and distributive processing development.

References

[1] I. Ramsen, Unsolved problems in chemistry, Modern Inventions and Discoveries, Hill, New York, 1904 (also quoted in H. Neurath: Why Protein Science? Protein Science 1 (1) (1992) 2).

[2] L.M. Gierasch, J. King (Eds.), Protein Folding American Association for the Advancement of Science, Washington, DC, 1990.

[3] M.A. Berg, G.A. Chasse, E. Deretey, A.K. Füzéry, B.M. Fung, D.Y.K. Fung, H. Henry-Riyad, A.C. Lin, M.L. Mak, A. Mantas, M. Patel, I.V. Repyakh, M. Staikova, S.J. Salpietro, T-H. Tang, J.C. Vank, A. Perczel, Ö Farkas, L.L. Torday, Z. Székely, I.G. Csizmadia, J. Mol. Struct. (Theochem) 500 (2000) 5 (Millennium Volume).

[4] I.G. Csizmadia, M.C. Harrison, B.T. Sutcliffe, Q. Prog. Rep. (MIT-SSMTG) 50 (1963) 1.

[5] I.G. Csizmadia, M.C. Harrison, B.T. Sutcliffe, Q. Prog. Rep. (MIT-SSMTG) 59 (1966) 43.

[6] I.G. Csizmadia, M.C. Harrison, B.T. Sutcliffe, Theor. Chim. Acta 6 (1966) 217.

[7] M.A. Robb, I.G. Csizmadia, Theor. Chim. Acta 10 (1968) 269.

[8] H.M. Basch, B. Robin, N.A. Kuebler, J. Chem. Phys. 47 (1967) 1201.

[9] I.A. Topol, S.K. Burt, E. Deretey, T.-H. Tang, A. Perczel, A. Rashin, I.G. Csizmadia, in preparation.

[10] M.J. Frisch, G.W. Trucks, H.B. Schlegel, P.M.W. Gill, B.G. Johnson, M.A. Robb, J.R. Cheeseman, T. Keith, G.A. Petersson, J.A. Montgomery, K. Raghavachari, M.A. Al-Laham, V.G. Zakrzewski, J.V. Ortiz, J.B. Foresman, J. Cioslowski, B.B. Stefanov, A. Nanayakkara, M. Challacombe, C.Y. Peng, P.Y. Ayala, W. Chen, M.W. Wong, J.L. Andres, E.S. Replogle, R. Gomperts, R.L. Martin, D.J. Fox, J.S.

Binkley, D.J. Defrees, J. Baker, J.P. Stewart, M. Head-Gordon, C. Gonzalez, J.A. Pople, Gaussian, Inc., Pittsburgh, PA, 1995.
[11] S.B. Prusiner, Philos. Trans. R. Soc. London B: Biol. Sci. 343 (1994) 447.
[12] G. Ivanovics, V. Bruckner, Naturwissenschaften 25 (1937) 250.
[13] G. Ivanovics, V. Bruckner, Z. Immunoforsch. 90 (1937) 304.
[14] J. Kovacs, V. Bruckner, Research 5 (1952) 194.
[15] J. Kovacs, V. Bruckner, J. Chem. Soc. (1952) 4255.
[16] J. Kovacs, V. Bruckner, K. Kovacs, J. Chem. Soc. (1953) 145.
[17] V. Bruckner, J. Kovacs, H. Nagy, J. Chem. Soc. (1953) 148.
[18] V. Bruckner, J. Kovacs, H. Nagy, Experientia 9 (1953) 63.
[19] V. Bruckner, J. Kovacs, K. Kovacs, J. Chem. Soc. (1953) 1512.
[20] V. Bruckner, J. Kovacs, K. Kovacs, Naturwissenschaften 39 (1952) 380.
[21] V. Bruckner, J. Kovacs, K. Kovacs, Naturwissenschaften 40 (1953) 243.
[22] V. Bruckner, J. Wein, M. Kajtar, K. Kovacs, Naturwissenschaften 42 (1955) 463.
[23] V. Bruckner, M. Szekerke, J. Kovacs, Naturwissenschaften 44 (1957) 90.
[24] V. Bruckner, J. Wein, M. Kajtar, K. Kovacs, Naturwissenschaften 44 (1957) 89.
[25] P.J. Flory, J. Chem. Phys. 10 (1942) 51.
[26] P.J. Flory, J. Chem. Phys. 17 (1949) 303.
[27] P.J. Flory, J. Chem. Phys. 18 (1950) 1086.
[28] G. Raos, G. Allegra, J. Chem. Phys. 107 (1997) 6479.
[29] F. Ganazzoli, G. Raos, G. Allegra, Macromol. Theor. Simul. 8 (1999) 65.
[30] G. Allegra, F. Ganazzoli, S. Bontempelli, Comput. Theor. Polym. Sci. 8 (1998) 209.
[31] I.M. Lifshitz, A.Y. Grosberg, A.R. Khokhlov, Rev. Mod. Phys. 50 (1978) 683.
[32] M. Doi, S.F. Edwards, The Theory of Polymer Dynamics, Clarendon Press, Oxford, 1986 (p. 27).
[33] S. Mizushima, T. Shimanouchi, S. Nagakura, K. Kuratani, M. Tsuboi, H. Baba, O. Fujioka, J. Am. Chem. Soc. 72 (1950) 3490.
[34] A.M. Buswell, J.R. Downing, W.H. Rodebush, J. Am. Chem. Soc. 62 (1940) 2759.
[35] R.E. Richards, H.W. Thompson, J. Chem. Soc. (1947) 1248.
[36] S.E. Darmon, G.B.B.M. Sutherland, Nature 164 (1949) 440.
[37] S.E. Darmon, G.B.B.M. Sutherland, J. Am. Chem. Soc. 69 (1947) 2074.
[38] S. Mizushima, T. Shimanouchi, M. Tsuboi, T. Sugita, K. Kurosaki, N. Mataga, R. Souda, J. Am. Chem. Soc. 74 (1952) 4639.
[39] J.J. Fox, A.E. Martin, Proc. R. Soc. A 162 (1937) 419.
[40] R.M. Badger, S.H. Bauer, J. Chem. Phys. 5 (1937) 839.
[41] M. Tsuboi, Bull. Chem. Soc. Jpn 22 (1949) 215.
[42] M. Tsuboi, Bull. Chem. Soc. Jpn 25 (1952) 385.
[43] G. Albrecht, R.B. Corey, J. Am. Chem. Soc. 61 (1939) 1087.
[44] H. Baba, A. Mukai, T. Shimanouchi, S. Mizushima, J. Chem. Soc. Jpn 70 (1949) 333.
[45] J.T. Edsall, J. Chem. Phys. 4 (1936) 1.
[46] J.T. Edsall, J. Phys. Chem. 41 (1937) 133.
[47] J.T. Edsall, J. Chem. Phys. 5 (1937) 225 (see also p. 508).
[48] J.T. Edsall, J. Am. Chem. Soc. 65 (1943) 1767.
[49] J.T. Edsall, H. Scheinberg, J. Chem. Phys. 8 (1940) 520.
[50] J.T. Edsall, J.W. Otvos, A. Rich, J. Am. Chem. Soc. 72 (1950) 474.
[51] I.M. Klotz, D.M. Gruen, J. Phys. Colloid Chem. 52 (1948) 961.
[52] M.M. Davies, G.B.B.M. Sutherland, J. Chem. Phys. 6 (1938) 755.
[53] N. Wright, J. Biol. Chem. 120 (1937) 641.
[54] N. Wright, J. Biol. Chem. 127 (1939) 137.
[55] S.E. Darmon, G.B.B.M. Sutherland, G.R. Tristram, Biochem. J. 42 (1948) 508.
[56] R.B. Corey, J. Donohue, J. Am. Chem. Soc. 72 (1950) 2899.
[57] K.H. Meyer, H. Mark, Chem. Ber. 61 (1928) 1932.
[58] W.T. Astbury, A. Street, Philos. Trans. A 230 (1931) 75.
[59] W.T. Astbury, H.J. Woods, Philos. Trans. A 232 (1933) 333.
[60] M.L. Huggins, Chem. Rev. 32 (1943) 195.
[61] T. Shimanouchi, S. Mizushima, Kagaku Sci. 17 (1947) 24 (p. 52).
[62] T. Shimanouchi, S. Mizushima, Bull. Chem. Soc. Jpn 21 (1948) 1.
[63] D. Crowfoot, Proc. R. Soc. London A 164 (1938) 580.
[64] W.T. Astbury, S. Dickinson, K. Bailey, Biochem. J. 24 (1935) 2351.
[65] H. Zahn, Z. Naturforsch. 2B (1947) 104.
[66] E.J. Ambrose, W.E. Hanby, Nature 163 (1949) 483.
[67] E.J. Ambrose, A. Elliott, R.B. Temple, Nature 163 (1949) 859.
[68] C.H. Bamford, W.E. Hanby, Nature 166 (1950) 829.
[69] W.T. Astbury, A. Street, Philos. Trans. R. Soc. 280 (1931) 75.
[70] L. Pauling, R.B. Corey, Proc. Natl. Acad. Sci. USA 37 (1951) 205 (see also pp. 235, 251, 256, 261, 272 and 282).
[71] M.F. Perutz, Nature 167 (1951) 1053.
[72] L. Bragg, J.C. Kendrew, M.F. Perutz, Proc. R. Soc. A 203 (1950) 321.
[73] G.S. Hartley, C. Robinson, Trans. Faraday Soc. 48 (1952) 847.
[74] J.C. Kendrew, R.E. Dickerson, B.E. Strandberg, R.G. Hart, D.R. Davies, D.C. Phillips, V.C. Shore, Nature 185 (1960) 442.
[75] M.F. Perutz, 63rd Harvey Lect. 63 (1969) 213.
[76] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, New York, 1986.
[77] R.R. Edelman, J.R. Hesselink, M.B. Zlatkin (Eds.), Clinical Magnetic Resonance Imaging 2nd ed., Saunders, Philadelphia, PA, 1996.
[78] H. Oschkinat, C. Griesinger, P.J. Kraulius, O.W. Sorenson, R.R. Ernst, A.M. Gronenborn, G.M. Clore, Nature 332 (1988) 374.
[79] G.W. Vuister, R. Boelens, R. Kaptein, J. Magn. Reson. 80 (1988) 176.
[80] O.A.L. El-Kabbani, E.B. Waygood, L.T.J. Delbaerel, J. Biol. Chem. 262 (1987) 1292.
[81] H. Roder, G.A. Elöve, S.W. Englander, Nature 335 (1988) 700.

[82] J.B. Udgaonkar, R.L. Baldwin, Nature 335 (1988) 694.

[83] T.F. Havel, K. Wüthrich, Bull. Math. Biol. 46 (1984) 673.

[84] W. Braun, N. Go, J. Mol. Biol. 186 (1985) 611.

[85] G. Wagner, W. Braun, T.F. Havel, T. Schauman, N. Go, K. Wurthrich, J. Mol. Biol. 196 (1987) 611.

[86] O.A.L. El-Kabbani, E.B. Waygood, L.T.J. Delbaerel, J. Biol. Chem. 262 (1987) 1292.

[87] P. Schultze, E. Worgotter, W. Braun, G. Wagner, M. Vasak, J.H.R. Kagi, K. Wurthrich, J. Mol. Biol. 203 (1998) 251.

[88] R.E. Klevit, B.E. Waygood, Biochemistry 25 (1986) 7774.

[89] J. Baum, C.M. Dobson, P.A. Evans, C. Hanlg, Biochemistry 28 (1989) 7.

[90] C. Anfinsen, The Molecular Basis of Evolution, Wiley, New York, 1959.

[91] S. Fraga, J.M.R. Parker, J.M. Pocock, Computer Simulations of Protein Structures and Interactions, Lecture Notes in Chemistry, vol. 66, Springer, Berlin, 1995 (XII + 282 pp.).

[92] P. De Santis, E. Giglio, A.M. Liquori, A. Ripamonti, Nature 206 (1965) 456.

[93] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, J. Mol. Biol. 7 (1963) 95.

[94] G.N. Ramachandran, Biopolymers 6 (1963) 1494.

[95] A.V. Guzzo, Biophys. J. 5 (1965) 809.

[96] P.Y. Chou, G. Fasman, Biochemistry 13 (1974) 211 (see also p. 222).

[97] P.Y. Chou, Abstracts, Second Chemical Congress of the North American Continent, 1980.

[98] D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Plenum Press, New York, 1989.

[99] A.T. Hagler, B. Honig, Proc. Natl. Acad. Sci. USA 75 (1978) 554.

[100] F.E. Cohen, M.J.E. Sternberg, J. Mol. Biol. 138 (1978) 321.

[101] S.B. Needleman, C.D. Wunsch, J. Mol. Biol. 48 (1970) 443.

[102] J. Kyte, R.F. Doolittle, J. Mol. Biol. 157 (1982) 105.

[103] B. Lee, F.M. Richards, J. Mol. Biol. 55 (1971) 379.

[104] G.M. Crippen, I.D. Kuntz, Int. J. Pept. Protein Res. 12 (1978) 47.

[105] B. Robson, E. Platt, R.V. Fishleigh, A. Marsden, P. Millard, J. Mol. Graphics 5 (1987) 8.

[106] G. Nemethy, H.A. Scheraga, Biopolymers 3 (1965) 155.

[107] A.M. Liquori, Q. Rev, Biophysics 2 (1969) 65.

[108] F.A. Momany, R.F. McGuire, A.W. Burgess, H.A. Scheraga, J. Phys. Chem. 79 (1975) 2361.

[109] A.T. Hagler, S. Lifson, P. Dauber, J. Am. Chem. Soc. 101 (1979) 5121.

[110] B.B. Brooks, R.E. Bruccelori, B.D. Olafson, D.J. States, S. Swaminatham, M. Karplus, J. Comput. Chem. 4 (1983) 187.

[111] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., P. Weiner, J. Am. Chem. Soc. 106 (1984) 765.

[112] S.J. Weiner, U.C. Singh, T.J. O'Donnel, P.A. Kollman, J. Am. Chem. Soc. 106 (1984) 6243.

[113] Y. Duan, P.A. Kollman, Science 282 (1998) 740.

[114] Y. Duan, L. Wang, P.A. Kollman, Proc. Natl. Acad. Sci. USA 95 (1998) 9897.

[115] C.L. Brooks III, M. Grueble, J.N. Onuchi, P.G. Wolynes, Chemical Physics of Protein Folding, Proc. Natl. Acad. Sci. USA 95 (1998) 11037–11038.

[116] A.M. Rodriguez, H.A. Baldoni, F. Suvive, R. Nieto-Vaquez, G. Zamarbide, R.D. Enriz, O. Farkas, A. Perczel, I.G. Csizmadia, J. Mol. Struct. (Theochem) 455 (1998) 275.

[117] M.A. McAllister, A. Perczel, P. Csaszar, W. Viviani, J.L. Rivail, I.G. Csizmadia, J. Mol. Struct. (Theochem) 288 (1993) 161.

[118] R. Hoffman, A. Imamura, Biopolymers 7 (1969) 207.

[119] A.J. Hopfinger, A.G. Walton, Bioplymers 9 (1970) 29.

[120] W. Damm, W.F. Van Gunsteren, J. Comp. Chem. 31 (2000) 774.

[121] J.F. Yan, F.A. Momany, R. Hoffman, H.A. Scheraga, J. Phys. Chem. 74 (1970) 420.

[122] F.A. Momany, R.F. McGuire, J.F. Yan, H.A. Scheraga, J. Phys. Chem. 75 (1971) 2286.

[123] D.A. Kleier, W.N. Lipscomb, Int. J. Quantum Chem., Quantum Biol. Symp. 4 (1977) 73.

[124] B. Maigret, B. Pullman, M. Dreyfus, J. Theor. Biol. 26 (1970) 321.

[125] B. Maigret, D. Perania, B. Pullman, J. Theor. Biol. 29 (1970) 275.

[126] B. Maigret, D. Perania, B. Pullman, Biopolymers 10 (1971) 491.

[127] R. Ramani, R.J. Boyd, Int. J. Quantum Chem., Quantum Biol. Symp. 8 (1981) 117.

[128] R. Ramani, R.J. Boyd, Can. J. Chem. 59 (1981) 3232.

[129] R.J. Boyd, J.S. Perkyns, R. Ramani, Can. J. Chem. 61 (1983) 1082.

[130] J.A. Ryan, J.L. Whitten, J. Am. Chem. Soc. 94 (1972) 2396.

[131] D. Peters, J. Peters, J. Mol. Struct. 53 (1979) 103.

[132] D. Peters, J. Peters, J. Mol. Struct. 109 (1984) 137.

[133] D. Peters, J. Peters, J. Mol. Struct. 68 (1980) 249.

[134] D. Peters, J. Peters, J. Mol. Struct. 69 (1980) 249.

[135] L.R. Wright, R.F. Boskman, J. Phys. Chem. 86 (1982) 3956.

[136] P.T. van Duijnen, B.T. Thole, Biopolymers 21 (1982) 1749.

[137] L. Schafer, C. van Alsenoy, J.N. Scardale, J. Chem. Phys. 76 (1982) 1439.

[138] J.N. Scardale, C. van Alsenoy, V.J. Klimkowski, L. Schafer, F.A. Momany, J. Am. Chem. Soc. 105 (1983) 3438.

[139] L. Schafer, V.J. Klimkowski, F.A. Momany, H. Chuman, C. van Alsenoy, Biopolymers 23 (1984) 2335.

[140] K. Sian, V.J. Klimkowski, C. van Alsenoy, J.D. Ewbank, L. Schafer, J. Mol. Struct. 152 (1987) 261.

[141] K. Sian, S.Q. Kulp, J.D. Ewbank, L. Schafer, C. van Alsenoy, J. Mol. Struct. (1989) 184.

[142] S.J. Weiner, V.C. Singh, T.J. O'Donnell, P.A. Kollman, J. Am. Chem. Soc. 106 (1984) 6243.

[143] A.M. Sapse, L.M. Fugler, D. Cowburn, Int. J. Quantum Chem. 29 (1986) 1241.

[144] T.C. Chean, S. Krimm, J. Mol. Struct. 188 (1989) 15.

[145] T.C. Chean, S. Krimm, J. Mol. Struct. 193 (1989) 1.

[146] A. Perczel, J.G. Angyan, M. Kajtar, W. Viviani, J.L. Rivail, J.F. Marcoccia, I.G. Csizmadia, J. Am. Chem. Soc. 113 (1991) 6256.

[147] T. Head-Gordon, M. Head-Gordon, M.J. Frish, C. Brooks,

J. Pople, Int. J. Quantum Chem, Quantum Biol. Symp. 16 (1989) 311.

[148] T. Head-Gordon, M. Head-Gordon, M.J. Frish, C. Brooks, J. Pople, J. Am. Chem. Soc. 113 (1991) 5989.

[149] R.F. Frey, J. Coffin, S.Q. Newton, M. Ramek, V.K.W. Cheng, F.A. Momany, L. Schafer, J. Am. Chem. Soc. 114 (1992) 5369.

[150] G. Endrédi, A. Perczel, Ö. Farkas, M.A. McAllister, G.I. Csonka, J. Ladik, I.G. Csizmadia, J. Mol. Struct. (Theochem) 391 (1997) 15.

[151] W. Viviani, J-L. Rivail, A. Perczel, I.G. Csizmadia, J. Am. Chem. Soc. 115 (1993) 8321.

[152] Ö. Farkas, M.A. McAllister, J.H. Ma, A. Perczel, M. Hollósi, I.G. Csizmadia, J. Mol. Struct. (Theochem) 369 (1996) 105.

[153] A. Perczel, Ö. Farkas, I.G. Csizmadia, Can. J. Chem. 75 (1997) 1120.

[154] Ö. Farkas, A. Perczel, J.F. Marcoccia, M. Hollósi, I.G. Csizmadia, J. Mol. Struct. (Theochem) 331 (1995) 27.

[155] A. Perczel, Ö. Farkas, I.G. Csizmadia, J. Comput. Chem. 17 (1996) 821.

[156] A. Perczel, Ö. Farkas, I.G. Csizmadia, J. Am. Chem. Soc. 118 (1996) 7809.

[157] H.A. Baldoni, A.M. Rodriguez, G. Zamarbide, R.D. Enriz, Ö. Farkas, P. Csaszar, L.L. Torday, C.P. Sosa, I. Jakli, A. Perczel, M. Hollosi, I.G. Csizmadia, J. Mol. Struct. (Theochem) 465 (1999) 79.

[158] S.J. Salpietro, A. Perczel, Ö. Farkas, R.D. Enriz, I.G. Csizmadia, J. Mol. Struct. (Theochem) 497 (2000) 39.

[159] M. Berg, S.J. Salpietro, I.G. Csizmadia, J. Mol. Struct. (Theochem) 504 (2000) 127.

[160] M.A. Zamora, H.A. Baldoni, A.M. Rodriguez, R.D. Enriz, C.P. Sosa, J.C. Vank, A. Perczel, A. Kucsman, Ö. Farkas, E. Deretey, I.G. Csizmadia, in preparation.

[161] J.C. Vank, C.P. Sosa, A. Perczel, I.G. Csizmadia, Can. J. Chem. 78 (2000) 395.

[162] O. Farkas, G.N. Zamarbide, H.A. Baldoni, L.L. Torday, A.M. Rodriguez, R.D. Enriz, C.P. Sosa, I. Jakli, A. Perczel, I.G. Csizmadia, Proline: The Maverick Amino Acid WATOC 99, 5th World Congress of Theoretically Oriented Chemists, Imperial College, London, UK, 1–6 August 1999, P212.

[163] K.D. Gibson, H.A. Scheraga, Biopolymers 4 (1966) 709.

[164] P.K. Ponnuswamy, V. Sasisekharan, Biopolymers 10 (1971) 565.

[165] G.N. Ramachandran, C.M. Vonkatacholam, S. Krimm, Biophys. J. 6 (1966) 849.

[166] P.N. Lewis, F.A. Momany, H.A. Scheraga, Isr. J. Chem. 11 (1973) 121.

[167] B. Pullman, A. Pullman, Adv. Protein Chem. 28 (1974) 34.

[168] A. Perczel, I.G. Csizmadia, Int. Rev. Phys. Chem. 14 (1995) 127.

[169] K. Kishikawa, F.A. Momany, H.A. Scheraga, Macromolecules 7 (1974) 797.

[170] M.A. McAllister, P. Csaszar, I.G. Csizmadia, J. Mol. Struct. (Theochem) 288 (1993) 181.

[171] A. Perczel, M.A. McAllister, P. Csaszar, I.G. Csizmadia, J. Am. Chem. Soc. 115 (1993) 4849.

[172] A. Perczel, M.A. McAllister, P. Csaszar, I.G. Csizmadia, Can. J. Chem. 72 (1994) 2050.

[173] M. Ramek, C.H. You, L. Schafer, Can. J. Chem. 76 (1998) 566.

[174] C.M. Liegener, G. Endrédi, M.A. McAllister, A. Perczel, J. Ladik, I.G. Csizmadia, J. Am. Chem. Soc. 115 (1993) 8275.

[175] G. Endrédi, C-M. Liegner, M.A. McAllister, A. Perczel, J. Ladik, I.G. Csizmadia, J. Mol. Struct. (Theochem) 306 (1994) 1.

[176] M. Cheung, M.E. McGovern, T. Jin, D.C. Zhao, M.A. McAllister, A. Perczel, P. Császár, I.G. Csizmadia, J. Mol. Struct. (Theochem) 309 (1994) 151.

[177] G. Endrédi, M.A. McAllister, A. Perczel, P. Császár, J. Ladik, I.G. Csizmadia, J. Mol. Struct. (Theochem) 331 (1995) 5.

[178] G. Endrédi, M.A. McAllister, Ö. Farkas, A. Perczel, J. Ladik, I.G. Csizmadia, J. Mol. Struct. (Theochem) 331 (1995) 11.

[179] L. Schafer, S.Q. Newton, M. Cao, A. Peters, C. Van Alsenog, K. Wolinsky, F.A. Momany, J. Am. Chem. Soc. 115 (1993) 272.

[180] A.G. Csázar, A. Perczel, Prog. Biophys. Mol. Biol. 71 (1999) 243.

[181] M.J. Böhm, S. Brode, J. Comput. Chem. 16 (1995) 146.

[182] C.P. Sosa, J. Ochterski, J. Carpenter, M.J. First, J. Comput. Chem. 19 (1998) 1053.

[183] A.D. Daniels, G.E. Scuseria, Ö. Farkas, H.B. Schlegel, Int. J. Quantum Chem. 77 (2000) 82.