Theoretical Investigation of Cooperative Effects of H-Bonds in Biomolecular Systems

vorgelegt von
Diplom-Chemiker
Rolf Martin Friedrich Streffer
aus Tübingen

Von der Fakultät II - Mathematik und Naturwissenschaften -
der Technischen Universität Berlin
zur Erlangung des Grades

Doktor der Naturwissenschaften
Dr. rer. nat.
genehmigte Dissertation

Promotionsausschuß:

Berichter: Prof. Dr. C.A. Dreismann
Berichter: Prof. Dr. G. Renger

Tag der mündlichen Prüfung: 21. Mai 2001

Berlin 2001
D83

Abstract

Streffer, Friedrich

Theoretical Investigation of Cooperative Effects of H-Bonds in Biomolecular Systems

The present work investigates signatures of cooperative effects of H-bonds especially in DNA with statistical and quantum chemical methods. Part I investigates results of information-theoretical analyses of DNA sequences of living organisms, which are tested qualitatively and quantitatively with respect to their biological aspects and/or implications. The results concern 'long-range correlations' and 'fractals' in intron-containing DNA sequences, their possible 'linguistic' structure, and other related aspects. The investigations demonstrate that the findings a 'fractal' structure in DNA are trivially equivalent to variations of the base pair composition, or patchiness, of different regions in a natural DNA sequence. It is explicitly shown that neither a well-defined 'scaling' or 'fractal' exponent, nor a well-defined Zipf exponent of the 'linguistic' test does exist. The biological origins of such variations are discussed. Quantitative comparisons of natural DNAs with computer-generated, artificial sequences are made. But the present work shows that certain natural DNA sequences (especially those with compact genomes) do have certain stochastic characteristics (say, pseudo-fractal exponents, averaged Zipf slopes, etc.) which are intrinsically different from artificial sequences. To shed more light on this point, investigations concerning the short range correlation of base pair, which may be caused by short-lived quantum entanglement of protons, are presented. The most striking finding is that quantum entanglement appears preferably between the third base of a codon and the first base of the following one. Results on a large number of DNA sequences of various types and from widely different taxa are reported. Additionally, results of current investigations concerning the so-called 'detrended fluctuation analysis' and the 'Kullback information measure' of DNA sequences are reported. Part II deals with quantum chemical investigations of the AT, GC and the artificial $\kappa\chi$ base pair. Different levels of theory have been applied with full geometry optimization and the 'frozen-core' approximation; among them are B3LYP/6-31G** and MP2/6-31G**. The calculations in the 'frozen-core' approximation confirm the double well character of all investigated potential energy surfaces, with a decreased energy of the transition states on the inclusion of the electron correlation effects. Couplings of the different proton transfer reactions are discussed, testing quantitavely the assumptions of a quantum mechanical Jordan-Block structure found in the GC pase pair. The geometry optimization of the relevant stationary points of the double proton transfer reactions in the AT and GC base pair has been performed at the B3LYP/6-31G** level of theory. The most striking finding is found during the normal mode analysis of the vibrations. A vibrational mode involving the relevant protons for the double proton transfer does exist in the GC base pair and all tautomers, whereas the AT base pair does not have such a vibrational mode. Implications are discussed. Finally, two experiments are proposed to test the findings.

Abstract

Streffer, Friedrich

Theoretische Untersuchungen von kooperativen Effekten der Wasserstoffbrückenbindungen in molekularbiologischen Systemen

Die vorliegende Arbeit untersucht Anzeichen, die durch kooperatives Verhalten von H-Bindungen im besonderen in der DNA verursacht wurden, mit statistischen und quantenchemischen Methoden.

Teil I beschäftigt sich unter anderem mit qualitativen und quantitativen Ergebnissen von Analysen aus der Informationstheorie, angewendet auf die DNA lebender Organismen. Die biologischen Aspekte und Implikationen werden betrachtet. Die Resultate betreffen sogenannte 'Korrelationen langer Reichweite' und 'Fraktale' in DNA Sequenzen, die Introns enthalten, und deren mögliche 'linguistische' Struktur, sowie weiter verwandte Aspekte. Die Untersuchungen zeigen, daß die o.g. Ergebnisse durch Variationen in der Basenzusammensetzung in verschiedenen Regionen einer natürlichen DNA hervorgerufen werden. Es wird explizit gezeigt, daß weder ein wohl definierter 'Skalierungs-' oder 'fraktaler' Exponent, noch ein wohl definierter Zipf Exponent des 'linguistischen' Testes existiert. Weiterhin zeigt Shannon's Redundanz Analyse keine besonderen Veränderungen bei einer 'Distanz' von 3 Basenpaaren für stark Protein kodierende Sequenzen, was aufgrund der Kodon-Struktur zu erwarten ist. Die biologischen Ursprünge solcher Variationen werden diskutiert. Natürliche DNA Sequenzen werden mit künstlichen, computer-generierten Sequenzen quantitativ verglichen. Es wird gezeigt, daß sehr wohl bestimmte natürliche DNA Sequenzen bestimmte stochastische Eigenschaften (z.B. pseudo-fraktale Exponenten, gemittelte Zipfgraphen usw.) aufzeigen, die sich intrinsisch von denen der künstlichen Sequenzen unterscheiden. Um dies näher zu betrachten, werden Untersuchungen präsentiert, die die Nahordung oder Korrelationen kurzer Reichweite betreffen, welche durch kurzlebige Quantenkorrelationen (Qantumentanglement) hervorgerufen werden. Dabei werden die Basenfolgen in 5'-3' Richtung AG, TG, TA, GC, CA, und CT als quantenkorreliert betrachtet und die Basenfolgen GA, GT, AT, CG, AC und TC als nicht quantenkorreliert. Das prägnanteste Ergebnis ist, daß diese Quantenkorrelationen vermehrt zwischen der dritten Base eines Kodons und der ersten Base des daraufolgenden Kodons auftreten. Es wird über Ergebnisse zahlreicher natürlicher DNA Sequenzen verschiedener Taxa berichtet. Zusätzlich werden Ergebnisse bezüglich der sogenannten 'detrended fluctuation analysis' und der Analyse von DNA Sequenzen mit Hilfe des Kullback Maßes der Informationstheorie gezeigt. Es zeigt sich, daß in der Abfolge der Basenpaare mehr Information enthalten ist als in der Abfolge der Pyrimidine und Purine.

Teil II befaßt sich mit der quantenchemischen Untersuchung der AT und GC Basenpaare, sowie des künstlichen $\kappa\chi$ Basenpaares. Dabei wurden verschiedene Grade der Näherung bei voller Geometrieoptimierung und bei fixierten Atomen, der sogenannten 'frozen-core' Näherung, angewandt. Unter den verwendeten Methoden befinden sich B3LYP/6-31G** und MP2/6-31G**. Die Untersuchungen mit der 'frozen-core' Näherung bestätigen den Doppel-Minimum Charakter der Hyperfläche. Der Energieabstand zwischen den Minima und dem Übergangszustand vermindert sich, wenn die Elektronenkorrelationsenergie mit berücksichtigt wird. Kopplungen der verschiedenen Protonentransferreaktionen werden diskutiert, und die Annahmen, die zu einer quantenmechanischen Jordan-Block Struktur im GC Basenpaar führen, werden quantitativ getestet und bestätigt. Die Jordan-Block Struktur führt dazu, daß zwei Freiheitsgrade, hier zwei Doppelprotonentransferreaktionen, zu einem Freiheitsgrad kolabieren, d.h. die Doppelprotonentransferreaktionen sind gekoppelt. Hierbei verläuft die relevante Differenz der Energie der Protonentransferreaktionen im GC Basenpaar linear für alle untersuchten Methoden.

Die Geometrieoptimierung der relevanten stationären Punkte der Protonentransferreaktionen wurde mit der B3LYP/6-31G** durchgeführt, welche für die kanonischen Basenpaare AT und GC eine exzellente Übereinstimmung mit den kristallographischen Daten für die Bindungslängen und Bindungswinkel liefert. Nach dieser Methode ist die pyramidale Konformation der oft diskutierten Aminogruppe des freien Adenin energetisch bevorzugt. Die Energiedifferenz zur planaren Konformation ist allerdings so gering, daß die Gruppe bei biologisch relevanten Temperaturen nicht in einer Konformation vorliegt. Es konnten keine stationären Punkte lokalisiert werden, die auf ionische Strukturen hinweisen. Weiterhin konnte im GC Basenpaar nur ein stationärer Punkt für ein Tautomer bestimmt werden. Ein stationärer Punkt für den Protonentransfer, an dem die H-Bindungen, die an der großen und kleinen Furche liegen, beteiligt sind, existiert nicht. Dies Ergebnis wird mit NMR Untersuchungen verglichen und diskutiert. Die relative Energie der Minima, die die Tautomere beschreiben, gegenüber den Übergangszuständen beträgt nach der Berücksichtigung der Nullpunktsschwingungsenergie -3.0 kcal mol^{-1} für das AT Basenpaar, bzw. -1.0 kcal mol^{-1} für das GC Basenpaar. Die in einem Doppelprotonentransfer beteiligten H-Bindungen werden um bis zu 0.4 Å verkürzt. Das prägnanteste Ergebnis wurde bei der Bestimmung der Normal-Moden der Schwingungen gefunden. Im GC Basenpaar und den Tautomeren existiert eine Schwingung an der alle relevanten Protonen des doppelprotonentransfers teilnehmen, während eine solche Schwingung im AT Basenpaar fehlt. Schlielich werden zwei Experimente vorgeschlagen, um einige Ergebisse dieser Arbeit zu testen.

Contents

1 Introduction

Hydrogen (H-)bonds appear in many physical, chemical and especially biological systems. This kind of bonding is of primary importance for the biomolecular systems in living organisms, since the latter always contain a significant number of H-bonding functional groups. This is surely related with the well-known facts that many biomolecules exhibit very specific H-bonding interactions with other molecules and that the evolution of life appeared in an aqueous medium. Moreover, the discovery of the double helical structure of DNA [1] showed that genetic information on our planet is only stored with H-bonds, using the Watson-Crick type patterns of H-bonds, which are essential for the structure and dynamics of DNA in living organisms.

There are many examples in biology of the association of ligands with a macromolecule being a cooperative process [2]. One of the most thoroughly studied examples is the association of oxygen with hemoglobin [3]. In addition, many multi-subunit enzymes bind substrates or other molecules in a cooperative fashion. E.g., the enzyme aspartatetranscarbamoylase exhibits this kind of behaviour [4]. Nucleic acids also bind, in different instances, particular ligands cooperatively.

In the past many chemists have thought only of electrons as quantum particles, which behave according to the laws of quantum mechanics, whereas nuclei behave according to classical mechanics the in potential 'created' by the electrons, due to their higher mass. This is often referred to as the 'Born-Oppenheimer' approximation. While the mentioned notion of the Born-Oppenheimer approximation is applicable for isolated molecules, i.e. in gas phase for most cases, there are experiments in which the quantum character of the nuclei is very important. Here one should especially mention the analogues of the Young double slit experiment with He [5] and Na [6] atoms as well as Na_2 [7] and C_{60} [8] molecules, in which the single atom or molecules are sent 'through a double slit' with a slit separation of several μm. The detector behind the 'double slit' detects an interference pattern proving the *single* particle to traverse through *both* slits at the same time. Therefore, the whole atom or molecule, with *all* its electrons and nuclei, has to be treated as *one* quantum system, which is delocalized across the two slits, i.e. the quantum system atom or molecule 'passes' through both slits at the same time.

Turning to condensed matter systems one may clearly detect the quantum nature of whole atoms at low temperature in chemical kinetics. The quantum nature of the involved protons shows up if the rate constant of a chemical reaction does not tend to 0, i.e no conversion at all, for low temperatures, but remains constant below a certain temperature. This is a characteristic behaviour of tunneling in the chemical reaction. But this method is restricted to very low temperatures.

In order to detect at least the tunneling of a single hydrogen nucleus under

physiological conditions, in the potential 'created' by its surroundings, isotope substitution of the hydrogen of interest by deuterium or tritium is employed. Hydrogen, deuterium and tritium have low but significantly different masses and therefore different de Broglie wavelengths λ_{dB}, which is a measure of how far quantum effects may extend, and the nuclei have different total spins. Especially, tunneling in chemical kinetics is studied by this approach, using the different λ_{dB} of H, D and T and therefore different tunneling frequencies in one and the same potential. So, it is possible to investigate the quantum character of the proton under consideration, i.e. the H which was exchanged by D or T, by following the global reaction kinetics without monitoring the microscopic process of interest itself.

There are several organic reactions in solution which exhibit at room temperature a kinetic behaviour characteristic for tunneling such as a H/D kinetic isotope effect in the order of 40 [9–11]. Under physiological conditions it is, furthermore, most interesting to know whether tunneling plays an important role for enzymatic reactions. At least for some enzymatic systems it could be shown that tunneling plays a significant role in the catalysed reaction. Amongst others there are alcohol dehydrogenase (ADH) [12], different amine oxidases [13–15], glucose oxidase [16], soybean lipoxygenase [17, 18], serine protease [19], lactate dehydrogenase [20], and carbonic anhydrase [21]. All the mentioned investigations look at the cleavage or formation of a C-H bond, since C-H bonds are known not to exchange the involved hydrogen isotope. Therefore an interpretation of the results in the frame of the so-called fractionation theory [22] is prevented.

The examples presented thus far have shown the quantum nature of single nuclei. When a single particle, i.e. a proton, shows quantum behaviour, it is legitimate to ask, if adjacent particles do form a single coherent quantum system. This could be the case for short times, when this quantum system is subject to the known decoherence effect, which is currently thought to be caused by the environment of the quantum system of interest [23].

Examples of systems having a very long decoherence time are the Bose-Einstein (BE) condensates at very low temperatures in which the atoms populate a single coherent quantum state [24]. E.g. it is possible to separate a cloud of such a condensate into two with some barrier [25]. Afterwards the two clouds may live for some while separated. In the following the barrier is switched off, and the clouds may extend to the space occupied by the other one. The clouds will interfere with each other, showing the characteristic interference pattern. This implicates that the two clouds still have a fixed phase relation. These experiments [24, 25] build the foundations of the atom laser. Again these systems are very isolated from their environment. Coupling the BE condensate to its surroundings by raising the temperature of interest above, say 100K, will always destroy it, by shortening the decoherence time. I.e. the coherence of the system will decay

faster, and the clouds will loose the ability to interfere with each other. Usually, it is assumed that at higher temperatures like room temperature, which are suitable for living organisms, any quantum coherence is totally smeared out or destroyed and is virtually not detectable by common experiments.

During the last years there have been considerable experimental efforts towards the detection of entangled states of protons (or deuterons) under moderate conditions. I.e. the temperature of interest is not some nano or micro Kelvin but in the range of room temperature ($\sim$300 K). As already mentioned above, with rising temperature the coherence time of the entangled states is shortened. Therefore the interaction time of the probe with the sample, i.e. the possibly entangled protons, has to be very short, more concrete, of the order of the decoherence time or shorter, otherwise one would not be able to detect any effect of the entangled protons (or deuterons). Furthermore, it is always much more difficult to measure absolute quantities than relative ones. Therefore and because quantum entanglement is supported by identical particles, in the following experiments always mixtures of protons and deuterons have been investigated. E.g. mixtures of water (H_2O) and heavy water (D_2O) whose composition is very well known by preparation (better than 0.1%). When quantum entanglement does play a role in the sample one expects a dependence of single particles properties on the sample composition, since the volume in which quantum entanglement does take place between identical particles, e.g. protons (or deuteron), is increasingly restricted the more particles of another kind are present.

A series of experiments started in 1995 with a Raman scattering experiment [26] which investigated a quantity R, the scattering of *one* O-H oscillator relative to *one* O-D oscillator, in water and heavy water mixtures of compositions ranging between the two limits pure water and pure heavy water. From the viewpoint of isolated oscillators (as one normally thinks of) the quantity R is expected to be a constant for all compositions ($R^{expected}$=1.97). In contrast to $R^{expected}$ the experiment finds a strong dependence of R on the composition of the mixtures investigated, e.g. R for a half-half mixture (x_D=0.5) is decreased by about 40%. The authors of [26] take this as an indirect indication of quantum entangled nuclei, since the light of the laser in the Raman experiment interacts with the electrons and the nuclei are only coupled to the electrons.

The Raman scattering experiment was followed by a Neutron Compton (or Deep Inelastic Neutron) scattering experiment [27], on water and heavy water mixtures, in which high energetic neutrons are scattered at the sample and the impulse and energy transfer are recorded. The major improvement over the Raman scattering experiment is that the neutrons interact directly with the nuclei (protons, deuterons etc.) and *not* with the electrons. Unfortunately the oxygen signal could not be resolved as a reference due to the experimental setup. Therefore, the ratio Q, the quotient of the total scattering cross-section of *one* proton and the total

scattering cross-section of *one* deuteron, was investigated. In the energy transfer regime of several eV (for this experiment 5 eV and larger) one expects $Q^{expected}$ to be 10.7, i.e. a proton scatters 10.7 times 'more' neutrons than a deuteron. This limits also the range of accessible mixtures to a minimal deuterium content of 30 % (x_D=0.3). Again the Neutron Compton Scattering (NCS) experiment finds a strong dependence of the ratio Q on the composition of the investigated mixtures, with the same tendency as the quantity R showed in the Raman scattering experiment. I.e. for low deuterium contents there is a significant decrease of the quantity under consideration (Q or R) which tends to the expected value $Q^{expected}$ and $R^{expected}$ respectively, for high deuterium contents. This is the first *direct* evidence for nuclear entanglement of light nuclei in condensed matter at ambient conditions.

Recently the NCS technique has been applied to urea solutions [28], metal hydrides (Nb-H-D and Pd-H-D) [29], mixtures of H_6-benzene and D_6-benzene as well as polystyrene with different deuterium contents. The results of all these experiments contradict the expectations of conventional theory as pointed out above. Two of these experiments are of particular interest, since their results hint at the connection to conventional theory.

The first of these two experiments is the NCS experiment on Nb-hydrides [29]. As compared with the H_2O-D_2O experiment the protons and deuterons may move between interstitial sites of the Nb lattice, but the Nb signal may be used as a reference, since no container enclosing the sample was used. Here it is necessary to mention that the scattering time or interaction time of the neutron with the struck nucleus is not a constant, but depends, among other parameters, on the momentum transfer during the scattering process [30], i.e. on the angle at which the scattered neutron is detected. In short, the higher the scattering angle, the shorter the scattering time. For the Nb-hydrides the scattering time ranges from about 0.1 to 1 femtosecond for the scattering of a neutron at a proton. Translating the scattering angle to scattering times for the protonic signal one finds a transition from non-conventional behaviour at short scattering times to conventional behavior at longer scattering times (> 0.7 fs). Furthermore this experiment gives for the first time an experimental estimate of the decoherence time of protons in a condensed matter system.

The second experiment is the NCS experiment on polystyrenes of different proton or deuterium content. In contrast to the Nb-hydrides there is, as in the experiments on H_2O-D_2O mixtures, no angular dependence of the proton or carbon signal, which is used as a reference. I.e. the decoherence time of the proton is different from the decoherence time of the protons in the Nb-hydride. But one finds a strong dependence of the proton signal on the deuterium content of the sample investigated. I.e. for the purely protonated sample one finds a reduction of the proton signal as compared with conventional expectations by 20%. As the pro-

tons are diluted progressively with deuterons the proton signal seems to increase and to approache the conventionally expected values.

Very recently, the electro chemical hydrogen evolution reaction (HER) in H_2O-D_2O mixtures has been investigated [31]. The experiments determine the reaction rate of the HER in different H_2O-D_2O mixtures, ranging from pure H_2O to pure D_2O, as a function of the applied potential. In contrast to the NCS experiments it is not possible to speak in terms of single particle properties, but also the reaction rate determined shows a pronounced deviation from the conventional expectations. I.e. assuming the validity of the fractionation theory [22] the results indicate that there are still 2.76 times more H^+ ions or 2.76 times less D^+ ions present in the molecular species reduced in a 50:50 mixture. The experimental results may be reproduced altering the most fundamental constant l[1] of fractionation theory to 0.25, in contrast to the so far established experimental value of l=0.69 [22, 32–34]. For a detailed discussion see Ref.[31].

In summary, common to all experimental results is a possible interpretation in the frame of effectivly altered particle numbers. I.e. during the experiment, if it is characterized by a sufficiently short time scale, a different number of particles seems to be present in the sample than during sample preparation and during the sample recheck after the experiment by a different method, e.g. densitometry for liquids. Affected are a wide range of hydrogen containing systems, not only systems with very mobile hydrogen like water or the metal-hydrides, but also systems containing more rigid hydrogen, e.g. C-H bonds in benzene or polystyrene. Furthermore, even though the quantum entanglement in condensed matter at ambient conditions lives only for short times (in order of femtoseconds [29]), chemical kinetics are altered on reducing the quantum entanglement, as shown in the HER [31].

Here a hole bunch of questions arise for biomolecular systems. First of all: Does cooperativity of protons introduced by quantum entanglement exist in biomolecular systems? The answer is certainly 'yes'. This is shown by the NCS experiments on benzene and especially the polystyrene which show that virtually every proton is affected on the relevant time scale. Furthermore, the β-sheet of polyglycine has been studied by Fillaux and coworkers [35]. The experiment showed a tunnel splitting of the protons connecting the backbones. This implies a perfectly symmetric double well potential seen by the proton between a nitrogen atom (N), to which the proton is covalently bonded, and an oxygen atom (O), which participates in the H-bond. This effect has no conventional explanation, but can be understood when introducing quantum entanglement between adjacent protons [36].

Second, does quantum entanglement or cooperativity of protons form some

[1] l is the equilibrium constant of the reaction $H + D^+ \Leftrightarrow H^+ + D$

'evolutionary advantage'? Searching for possibilities to handle this question, the basic work of Eschenmoser and co-workers [37, 38] about novel structural forms of nucleic acids (i.e., Homo-DNA and P-RNA) and their possible consequences for prebiotic biomolecular evolution has also been noted. The secondary structures of the novel Homo-DNA form does not exhibit the well-known helical structure of the natural B-form DNA [37, 38]. In the light of these results, the preceding speculative question can be extended and/or more precisely formulated in the following way: Could it be that a related quantum-correlation effect also affects H-bonds belonging to base pairs of B-form DNA?

The crucial point now is: this question is not any more esoteric, since in principle it can be tested 'experimentally', i.e. it can be confronted with experimental data. Namely, if one assumes that the considered quantum-correlation effect does really cause some kind of evolutionary advantage or disadvantage (however 'small' it may be), then one would expect that Nature has already made some 'use' of it after about three billion years of natural selection and/or evolution of DNA molecules. But then this 'use' could be manifested in specific 'features' of, or 'patterns' appearing in, DNA nucleotide sequences of present-day living organisms, a large number of which being already available in DNA data bases.

In part I various statistical methods are investigated to unveil cooperativity of two (or more) H-bonds, which has been manifested in certain patterns in DNA even at large distances, e.g. some hundreds or thousands of base pairs [39]. The second part deals with a more meso- and microscopic view of cooperativity. Here the coupling of different quantum mechanical degrees of freedom is investigated. In the outlook some experiments will be proposed to test some of the findings.

1.1 A theoretical glimpse at quantum entanglement

Let S_A and S_B be two quantum systems associated with the density operators ρ_A and ρ_B and the associated Hilbert spaces H_A and H_B, which are spanned by the state vectors $|a_i\rangle$ and $|b_i\rangle$. If the quantum systems S_A and S_B interact, or if they have interacted in the past, with each other, then it is well known that the total wave function describing both systems exhibits entanglement[2]. In that case, the density operator ρ_{A+B} of the complete system has not the simple product form $\rho_A \otimes \rho_B$, but it may be given by

$$\rho_{A+B} = \sum_{k,l} c_k c_l^* |a_k b_k\rangle\langle a_l b_l| \tag{1.1}$$

The complex numbers c_k represent probability amplitudes. Standard quantum theory implies that measurements on an ensemble of S_A systems are described

[2] 'Pathological' forms of interaction Hamiltonians and associated possible exceptions from this generally valid rule are not considered here.

with the aid of the reduced density operator

$$\rho_A \equiv \mathrm{Tr}_B(\rho_{A+B}) = \sum_{k,l} |c_k|^2 |a_k\rangle\langle a_k| \tag{1.2}$$

(Tr_B denotes the partial trace over the variables of S_B.) A similar equation holds also for ρ_B. As a result, the maximal physical information being accessible by single measurements on the subsystems S_A and S_B is given by

$$\rho_{nc} = \rho_A \otimes \rho_B \tag{1.3}$$

where the subscript 'nc' refers to (ensembles of) non-correlated systems. Comparison with Eq. 1.1 shows that the quantal phase factors between the (generally complex) amplitudes c_k are completely lost in the case of Eq. 1.3. Thus the physical information one can obtain from all possible measurements on each of the two subsystems does not allow the reconstruction of the complete density operator ρ_{A+B} representing ensembles of entangled systems, since in general holds

$$\rho_{A+B} \neq \rho_{nc} \tag{1.4}$$

The aforementioned phase factors between the amplitudes c_k represent physically the quantum correlations, or entanglement, between the subsystems S_A and S_B. These phase factors can be determined only by appropriate measurements on *both* subsystems *at the same time*. In that case, one says the corresponding experiments measure the quantum correlations (or the degree of entanglement) between the subsystems.

As is obvious from Eq. 1.1, quantum correlations are mathematically represented by the off-diagonal matrix elements of the density operator ρ_{A+B}. The well known coherent superposition of states refer to a specific kind of quantum correlation, which – illustratively speaking – may be regarded as the maximal, or the most intense, form of entanglement.

Relating the presented Raman [26], NCS [27–29] and HER [31] experiments to quantum entanglement leads to a symmetry consideration of a matrix element of the type

$$M = \langle \psi(probe)\Phi(H,H)|H_{int}|\psi_{ex}(probe,H)\psi_{gr}(H)\rangle \tag{1.5}$$

which is, when squared, proportional to the scattered light or neutron field, in case of the Raman or NCS experiment, and proportional to the current density, in the HER experiment, of a pure H_2O sample. $\psi(probe)$ denotes the wave function of the probe, e.g. the light quanta, the neutron, etc., before the interaction with the sample. $\Phi(H,H)$ represents the quantum entangled wave function of two protons before the interaction. Note, $\Phi(H,H)$ is not separable, i.e. it may *not* be

written as a product state. Furthermore, the wave function of the proton measured and the probe after the interaction, which may be a product state, is described by $\psi_{ex}(probe, H)$ and $\psi_{gr}(H)$ is the wave function of the proton which is 'untouched' by the probe.

Remember, due to the spin superselection rule of quantum theory, the quantum entanglement between protons (and deuterons) is expected to be suppressed in isotopic mixtures of protons and deuterons. It is assumed that quantum entanglement between protons, which are fermions, in pure H_2O at ambient conditions is disturbed by the presence of deuterons, which are bosons and vice versa, giving rise to the observed deviations.

The picture of quantum entanglement and decoherence in condensed matter at ambient conditions given, is by no means complete, nore does a complete theory exist. But at least some characteristics of quantum entanglement in condensed matter have been established experimentally. So the distance between the entangled particles, i.e. a few Angstroms, and the time scale of vibrations are relevant. Both parameters are also important in biomolecular systems, which are especially famous for their ability to optimize their geometry to their needs. Some examples are the structure of DNA, which protects intrinsically the H-bonds encoding the genetic information, or the enzymes which have optimized their geometry to stabilize transition states of different reactions.

Part I
Statistical Analyses of DNA

2 Introduction

The structure of biological macromolecules like proteins, RNA and DNA, and to some extent also their dynamics, has always fascinated not only biologists and chemists, but physicists and mathematicians, too. In particular, after the discovery of the double helical structure of DNA by Watson and Crick in the year 1953 [1], this molecule has often been considered by physicists as an 'aperiodic solid', and many efforts have been undertaken in order to 'understand' its structure in the framework of solid state theory and/or quantum chemistry; see [40]. About ten years ago, for instance, one claimed the experimental observation of soliton-like energy propagation in DNA, as well as its physical interpretation. These findings, however, have been clearly disproved by detailed experimental investigations at the universities of Uppsala and Oxford [41, 42]. The universality of the genetic code and the geometric structure of proteins attracted the attention of many theoreticians.

During the last years, several reports have described the occurrence of a special kind of *long-range correlations* between nucleotides in DNA sequences [39, 43–45]; see also [46]. Of particular interest was the claim that such correlations appear only in intron-containing sequences [39, 43, 44], whereas intron-less sequences show no long-range correlations, whether of eukaryotic or prokaryotic origin. According to certain related results by Voss [45], which however do not distinguish between intron-containing and intron-less sequences, the considered correlations were proposed to differ between evolutionary categories of organisms.

It is obvious that the claimed observations of long-range correlations could be of far reaching nature [47–49]. These correlations, if firmly established, could also be of physical as well as chemical interest, since long-range correlations are often considered to be associated with the existence of certain specific non-equilibrium dynamically processes, fractals and/or chaotic structures, etc.; see the original references [39, 45]. In this sense, it was mentioned by Peng et al. [39] that intron-less sequences seem to exist in an equilibrium state (with maximum entropy), whereas intron-containing sequences might be far from thermodynamic equilibrium. It has even been speculated that these findings may shed light upon the biological significance of introns [39, 48]. However, another (more down-to-earth) proposition was that the long-range correlations in intron-containing sequences may be plainly caused by repeated segments in introns; cf. [44, 50].

Newer investigations have questioned the claimed difference between intron-containing and intron-less DNA sequences [51–59] and, moreover, they pointed out that the 'long-range correlations' seem to appear whenever relatively large variations in nucleotide composition along the DNA sequence are present [53–57]. With the aid of examples it was demonstrated that these variations have clear biological origins [58]. In particular, investigations based on computer simulations [56, 57] demonstrated unequivocally the intrinsic connection between compositional heterogeneity or patchiness [53] of a DNA sequence and the appearance of claimed [39] long-range correlations and/or fractality. These topics are discussed in Section 3.

In Section 5 the recent findings of Mantegna et al. [60] concerning the *'linguistic'* features of natural DNA, and especially of non-coding DNA sequences, are considered in some detail . Certain statistical investigations, called 'linguistic tests', on DNA sequences have been performed, which are related to Zipf's distribution [61] and Shannon's information theory and redundancy analysis [62, 63]. These tests [60] are claimed to reveal new and significant differences between coding and non-coding parts of natural DNA sequences. The first of these linguistic tests is related to the so-called Zipf plot, i.e. the relation between the relative occurrence of all oligonucleotides of a given length n (called 'words' with n letters) in a specific DNA sequence (called a 'text'). Among other points, it was claimed [60] that:

- in a double-logarithmic plot, the graphs of the aforementioned relation for different DNA sequences are linear, which implies that Zipf's law [61] applies to the present case;

- the slopes of these graphs for coding and non-coding DNA sequences differ significantly.

The second linguistic test uses the information theoretical 'entropy' $H(n)$ [62, 63] of a DNA sequence (i.e. a 'text') when it is viewed as a collection of n-tuple words, as well as the associated 'redundancy' defined by Shannon; see below. This 'redundancy' may be considered as a property of natural languages, the purpose of which being to preserve the meaning of a text also in the case of 'typographical errors'.

As a result of these investigations, it was claimed [60] that - in clear contrast to protein coding DNA segments – the non-coding DNA parts are related with a considerable amount of redundancy, which corresponds to 'another sign that something was written in these mysterious stretches' [64].

Very recently, however, these findings and/or claims have been strongly criticized by Konopka and Martindale [65] by stressing, among others, the following points:

- Statistical differences of coding and non-coding DNA are known at least since 1981, which are used even in routine methods for discrimination between them; therefore, the claimed novelty of the results was not appreciated.

- The oligonucleotide frequency distribution in non-coding DNA does not appear to fit Zipf's law any better than does the distribution in coding regions; additionally the presented log-log plots display a non-linear rather than a linear trend.

- It was concluded [65] that both coding and non-coding DNA regions fit Zipf's law rather poorly, if at all.

The aforementioned findings and/or claims [60] are interesting and could have potentially a thus far unknown biological significance. Therefore they are investigated in more detail. In order to make the contribution as precise and as possible, it concentrates on a quantitative analysis which is mainly based on computer simulated sequences and their comparison with natural DNA sequences. In other words, a conceptually similar procedure as in the tests of 'long-range correlations' is applied. The results of these analyses, seem not to support the aforementioned claims [60].

In section 4 the so-called 'detrended fluctuation analysis' [66, 67] is shortly considered, which is a newly presented, and probably improved, method for the detection of long-range correlations in patchy DNA sequences. In section 6, some possible biological origins for the appearance of 'long-range correlations' are presented. This section, in contrast to the other main sections of this part, deals with molecular-biological features of specific natural DNA sequences. In section 7 some results of an analysis of DNA sequences based on the information theoretical Kullback measure [59, 68, 69] are presented. Section 8 concerns short range ordering of base pairs based on the ideas of quantum entanglemant of H-bonds.

3 On 'Fractality' of DNA

3.1 Theoretical remarks

3.1.1 The 'fractal' or 'pseudo-fractal' exponent α

Here the underlying method of the 'scaling' or 'fractal' exponent associated with a 'DNA random walk' [39] and the explicit numerical procedure [56, 57] is described, which permits the quantitative comparison of natural DNA sequences with artificial ones.

The 'scaling' exponent of a DNA sequence is defined as follows: Firstly, a new function $u(i)$ is defined, putting $u(i) = +1$ (or $= -1$) for a pyrimidine (or a purine) at the i-th position of the sequence. Secondly, the partial sum is defined

$$\begin{aligned} y(k) &= \sum_{x=1}^{k} u(x) \\ \Delta y(k, k_0) &= y(k_0 + k) - y(k_0) \end{aligned} \tag{3.1}$$

which simply gives the number of pyrimidines minus that of purines in a sequence interval of length k, starting at base pair position (k_0+1). This partial sum depends on the number k_0, i.e. it is a function of k_0. Thirdly, the variance of $\Delta y(k, k_0)$ is calculated,

$$\begin{aligned} F^2(k) &= \langle \Delta y(k, k_0)^2 \rangle - \langle \Delta y(k, k_0) \rangle^2 \\ &\equiv const \times k^{2\alpha^*(k)} \end{aligned} \tag{3.2}$$

In almost all cases studied $const \approx 1$ holds true.

The brackets represent an average over all possible positions k_0 of a base in the sequence. The identity sign ($\equiv$) stresses the fact that the second equation is tantamount to the definition of the new function $\alpha^*(k)$. In other words, the exponent $\alpha^*(k)$, for every value of k, is just a fitting parameter to the data given by $F^2(k)$. k is sometimes called the 'correlation length'. If and only if $\alpha^*(k)$ takes a constant value for all values of k, the existence of a 'scaling' or 'fractal' exponent may be assumed characterizing the DNA sequence under consideration. Contrary to the above claims [39], however, it has been shown by several authors [51–53] that, in all tested DNAs, the graph of $\alpha^*(k)$ is strongly curved, and thus, there does not exist a well-defined scaling exponent and the associated 'fractal' structure. As an approximation, let α be simply defined as the average over all

$\alpha^*(k)$ values

$$\alpha = \frac{1}{k_{max} - 1} \sum_{2}^{k_{max}} \alpha^*(k) \tag{3.3}$$

where k_{max} is the chosen maximum value of k. According to ref. [39], k_{max} should not exceed one tenth of the length of the DNA sequence. Simply speaking, this α is just a fitting parameter and might be called 'pseudo scaling' or 'pseudo-fractal' exponent.

3.1.2 The local 'pseudo-fractal' exponent $\alpha(i)$

In order to investigate a sequence in more detail the local 'pseudo-fractal' exponent was introduced in ref. [56]. This method focuses on subsequences of a DNA sequence with an appropriate but constant length L_i centered at position i of the DNA sequence, and calculates the 'pseudo- fractal' exponent α_{local} over this subsequence. (L_i often equals 5000 bp, since it produces relatively 'good statistics' and it is not 'too long'.) Thus α_{local} is a local property of the complete DNA sequence 'at position i', and it is written $\alpha(i) = \alpha_{local}$. To obtain $\alpha(i)$ along the entire sequence (i.e., for all possible values of i), i is moved through the sequence with some appropriate step (typically between 200 and 500 bp).

3.1.3 Generation of artificial sequences

Here the generating procedure of an artificial sequence associated to a natural DNA strand [56] is described. This means the artificial sequence will have the same length and a very similar base composition.

1. Choose an arbitrary, but constant, interval D_i (often D_i equal 100 bp)

2. Divide the natural DNA-sequence in subsequences of length D_i and calculate the ratio $R(i)$ of pyrimidines/purines (or any other base combination, e.g. A,T/C,G) in each subsequence (Instead of the notation A,T/C,G,it may also equivalently be written in short: AT/CG). Clearly, the value of $R(i)$ may be considered as a 'fingerprint' of the genomic base composition around the i-th position of the DNA sequence. In the calculation of the $R(i)$ value of a (sub-)sequence the number of pyrimidines and purines are counted separately and divided:

$$R(i) = \frac{pyrimidines}{purines} \tag{3.4}$$

3. Generate for each $R(i)$ data point with an appropriate computer program, which creates a random series of pyrimidines and purines, a base sequence of length D_i. The deviation in the base composition ($R(i)$ value) is within the statistical error, depending on the quality of the used random number generator, which is the main part of the program.

4. Concatenate all randomly generated base strings in order to create an artificial sequence of the same length as the DNA sequence. This artificial sequence has then similar pyrimidines/purines ratios along its length to the original DNA sequence.

Numerical investigations showed, that the base composition of the natural DNA and the artificial sequence do not differ significantly. The base composition in artificial sequence associated to a 50000 bp long DNA differs only by about 40 bases, which corresponds to 0.08%.

3.1.4 Deviation between DNA sequences and associated artificial sequences $\Delta\alpha$

During the investigation of DNA sequences it turns out to be necessary to compare DNAs with their associated artificial sequences in a quantitative way. Therefore the quantity $\Delta\alpha$ has been introduced by us in ref. [57]. $\Delta\alpha$ is a measure of the mean deviation of the statistical behaviour (represented by the local 'pseudo-fractal' exponent $\alpha(i)$) of a natural DNA and the associated artificial sequences.

The first step in the calculation of $\Delta\alpha$ is to obtain the local 'pseudo-fractal' exponent functions of the natural DNA sequence $\alpha_{DNA}(i)$, and of a number of associated artificial sequences denoted by $\alpha_{art}(i)$. Secondly, the deviation $\delta\alpha$ of one artificial sequence and the DNA is calculated:

$$\delta\alpha = \frac{1}{M}\sum_{i=1}^{M}[\alpha_{DNA}(i) - \alpha_{art}(i)] \tag{3.5}$$

M counts all possible positions i. M depends on the parameters in the determination of the local 'pseudo-fractal' exponents. To obtain the quantity $\Delta\alpha$, $\delta\alpha$ is averaged over all individual values (denoted by the brackets):

$$\Delta\alpha = \frac{\langle\delta\alpha\rangle}{\alpha_{DNA}} \cdot 100\% \tag{3.6}$$

α_{DNA} is equal to the arithmetic average of all values of $\alpha_{DNA}(i)$. In most calculations 10 different associated artificial sequences are used to derive $\Delta\alpha$.

With this procedure the following general problem in the comparison of sequences is avoided. In the derivation of the 'pseudo-fractal' exponent of a random

sequence ($\alpha = 0.5$) a constant base composition is assumed (corresponding to a constant ratio $R(i)$). But in fact all DNA sequences have a varying base composition along the sequence, and our numerical procedure for the creation of associated artificial sequences takes this into account. As a consequence, $\alpha = 0.5$ is not characteristic for these artificial sequences.

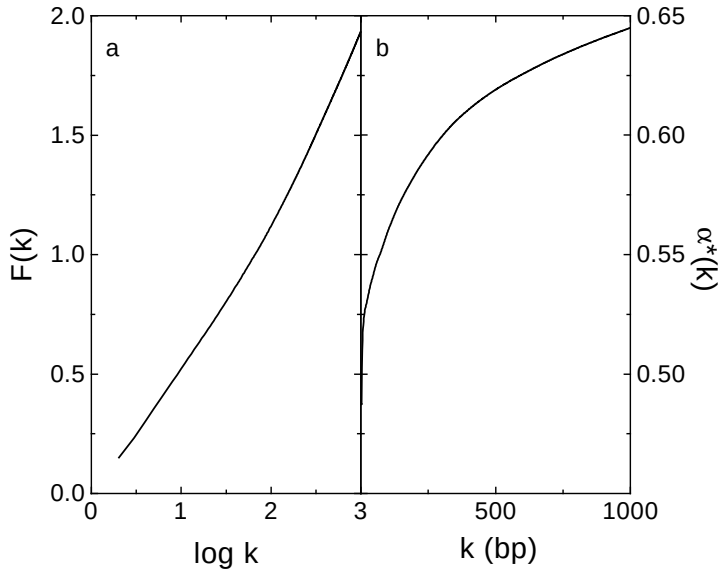

Figure 3.1: The 'optical deception' of the double logarithmic plot used by Peng et al. [39] is demonstrated. The genome of bacteriophage λ is chosen, since it shows clearly no straight line, see (a). (a) displays the double logarithmic plot fluctuation $F(k)$ versus the correlation length k. The slope of the graph would decide on the 'pseudo fractal' exponent α (b) shows the function $\alpha^*(k)$, which has a strong curvature and not a constant.

3.2 Results

3.2.1 Linearity of the fluctuation $F(k)$ in double logarithmic plots

The main result of Peng et al. in ref. [39] is the claim of long-range correlations and the associated 'fractality' of DNA sequences. This interpretation is based on the power law dependence of the fluctuation $F(k)$, i.e. with $\alpha^*(k) \equiv \alpha = const$ for all k. One possible test of this behaviour, used in ref. [39], is to plot the fluctuation $F(k)$ against the correlation length k on double logarithmic paper, which has to show a linear graph if the DNA sequence under consideration fits the mentioned condition. In fig. 3.1a an analysis of the λ-phage genome is presented in this manner. On first sight, it seems to be fairly linear for about two and a half

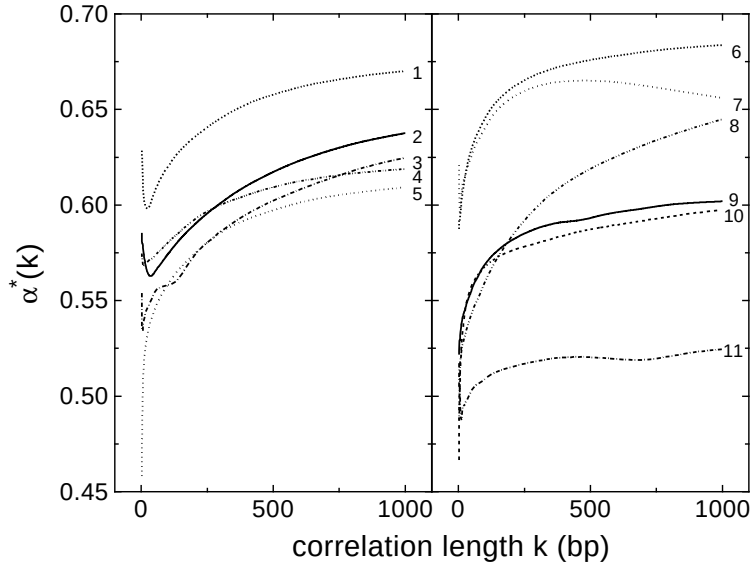

Figure 3.2: Some typical graphs of the 'pseudo fractal' exponents $\alpha^*(k)$ for complete DNA sequences. Contrary to the claims of Peng et al. [39], all these graphs are strongly curved, explicitly the $\alpha^*(k)$ graphs of the following sequences are shown: 1 *Paramecium Aurelia* mitochondrion (MIPAGEN) 2. *Marchantia Polymorpha* chloroplast (CHMPXX) 3. *Homo Sapiens* mitochondrion (MIHSXX) 4. *Marchantia Polymorpha* mitochondrion (MIMPCG) 5. *Saccharomyces Cerevisiae* [yeast] mitochondrion (MISCCG) 6. human β-globin region on chromosome 11 (HSHBB) 7. human adenosin deaminase (HSADAG) 8. bacteriophage λ (LAMBDA) 9. adenovirus type 2 (AD2) 10. human *cytomegalovirus* strain AD169 (HEHCMVCG) 11. bacteriophage T7 (PODOT7). Acronyms in parentheses denote the EMBL database identification name of each sequence. All calculations are performed with a maximum correlation length k_{max} =1000.

decades. But is it really linear? As mentioned in the introduction this could have far reaching consequences. So the function $\alpha^*(k)$ was derived with a numerical differentiation (to be more precise: with a method approximating the numerical differentiation) and plotted against the correlation length k [52]. If the double logarithmic plot shows a linear graph indeed the function $\alpha^*(k)$ has to be a constant or, more realistic, $\alpha^*(k)$ may scatter around a mean value. fig. 3.1b shows explicitly that $\alpha^*(k)$ is neither a constant nor scattering around some mean value. Instead, $\alpha^*(k)$ is monotonously growing, proving the graph in double logarithmic plot of fig. 3.1a to be curved [51–53, 56–58]. This behaviour is not restricted to the genome of the bacteriophage λ, but is found in all DNA sequences investigated thus far (see fig. 3.2 for some examples).

Furthermore, Karlin and Brendel [53] concluded that the pervasive patchiness of DNA usually precludes modeling DNA sequences by a mathematical stationary process, and, in particular, by such a process having long-range dependence. Moreover, due to compositional heterogeneity of long DNA sequences (and especially for those containing introns), attempts to break up the sequence into more homogeneous segments (as Peng et al. [39] did) cannot succeed, unless the segments are chosen to be so small that the whole notion of long-range correlations evaporates [53].

3.2.2 Differences between intron-less and intron-containing sequences

Another claim of Peng et al. [39] concerns the distinction in the magnitude of α between intronless and intron-containing DNA sequences. This finding aswell could not be confirmed by several groups [45–47, 57]. The pseudo-fractal parameters α presented in table 3.1 (see the Appendix), which characterize the various *complete* DNA sequences under consideration, do not exhibit significant differences of the claimed kind. This is also supported by the results of a study of the variations of the local 'pseudo-fractal' exponent $\alpha(i)$ along a DNA sequence.

$$\alpha = \frac{1}{M} \sum_{i=1}^{M} \alpha(i) \tag{3.7}$$

In order to prevent possible confusion, let us stress that α is approximately the average over all $\alpha(i)$ values, where the index i runs along the sequence. The local 'pseudo-fractal' exponent $\alpha(i)$ varies strongly along all DNA sequences. Two striking examples are displayed in fig. 3.6. Moreover, and more importantly, the variations within a sequence are in nearly all cases larger than the sequence-to-sequence differences between the pseudo-fractal parameters α.

3.2.3 'Long-range' correlation evaporates

The third point concerns the origin of α values around 0.6, interpreted as fractal behaviour of the sequences [39] under consideration. The analysis is based, first of all, on the definition of the function $R(i)$ (definition of $R(i)$ is given in section 3.1.3 step (2)), which represents the pyrimidine/purine ratio around base positions i of the DNA sequence.

As an example the ratio $R(i)$ for the bacteriophage λ genome is presented in fig. 3.3a. There are three distinct regions (I, II and III; cf. also reference [70] for a description of these regions), in which $R(i)$ appears to fluctuate randomly around different constant values (In region III, $R(i)$ seems to increase with a constant slope). As shown in fig. 3.3, the mean values α of the exponent $\alpha^*(k)$

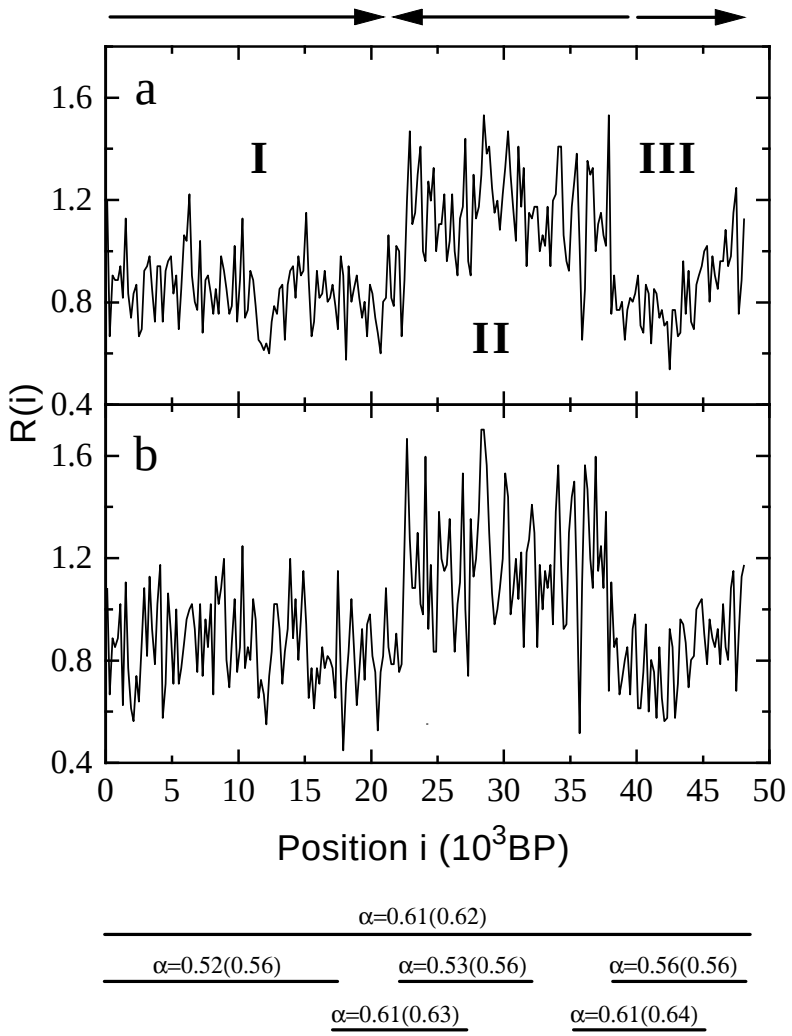

Figure 3.3: The ratio $R(i)$ of pyrimidines/purines of the λ phage DNA sequence (a), and $R(i)$ of a typical artificial sequence (b), with regions I, II and III as mentioned in the text. The α values given below the graphs are calculated over the regions indicated by the bars. Numbers in parentheses refer to the artificial (i.e. computer generated) sequence. The arrows indicate the direction of transcription.

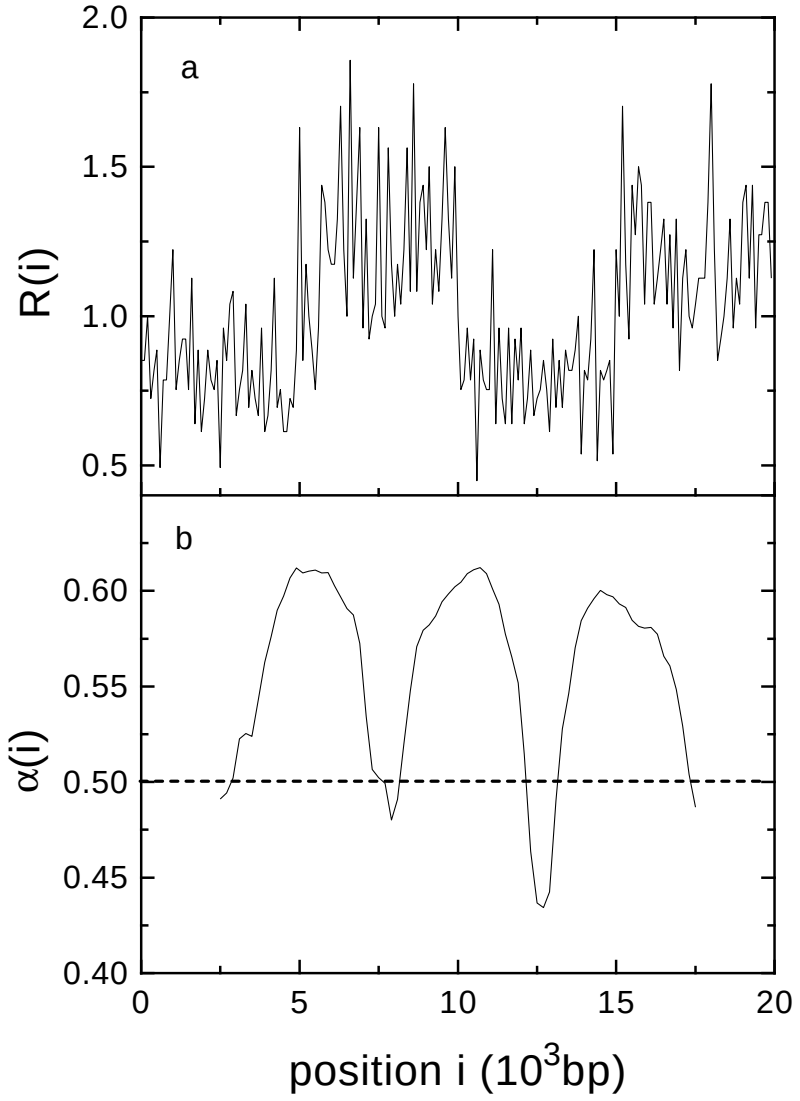

Figure 3.4: Dependence of the $\alpha(i)$ graph of a periodic artificial sequence on the periodicity (equal to $2D$) of the pyrimidine/purine ratio as shown in the $R(i)$ graph of this sequence. (a) Graph of $R(i)$, here with D =5000 'base pairs'. $R(i)$ varies between the mean values 0.8 and 1.2 (cf. the text). (b) The α(i) graph of this sequence. In this and the following figs. of this section the parameters $L_i =$ 5000 and $k_{max} = 500$ were used.

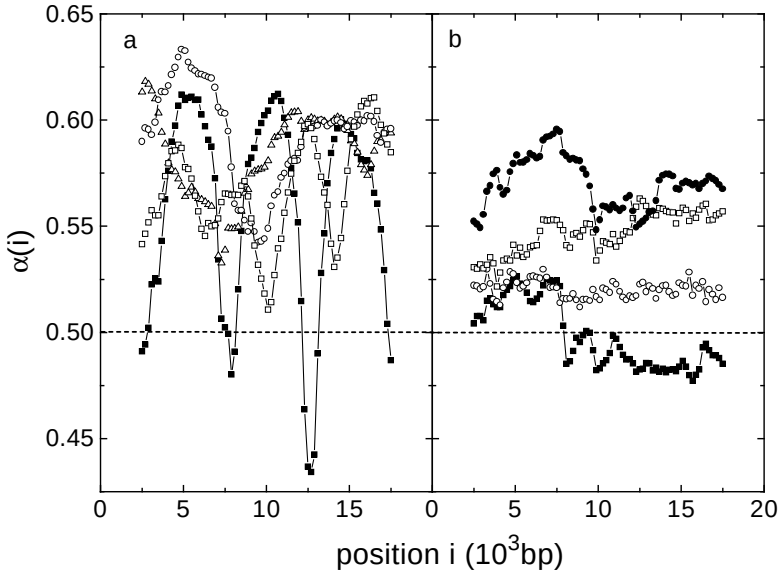

Figure 3.5: $\alpha(i)$ graphs of different periodic artificial sequences and their dependence on the periodicity of the pyrimidine/purine ratio. (a) ■ : $D = 5000$; □ : $D = 4000$; ∘ : $D = 2000$; △ : $D = 1000$. Note the disappearance of the periodicities with decreasing value of D . The average value of α is about 0.58, which mimics a 'fractal' sequence structure. (b) • : $D = 500$; □ : $D = 300$; ∘ : $D = 200$; ■ : $D = 100$. Note that the average value of α decreases continuously from 0.56 to the 'non-fractal' limit of 0.5 for the smallest values of D.

calculated within each of these regions yield values near to the 'random' limit α=0.5, which indicates no 'fractal' character of the sequence [39]. On the other hand, calculating $\alpha^*(k)$ over the complete sequence result in an average value of α = 0.61, which indicates a 'fractal' character of the sequence [39]. Moreover, in subsequence of the bacteriophage λ genome with a significantly changing ratio $R(i)$, the calculated large α-values indicate a 'fractal' character of the sequence, too (see fig. 3.3 and ref. [58]).

These observations are obviously contradictory, and therefore they motivated the comparison of the DNA sequences with associated artificial sequences being produced with the aid of random number generators (cf. Section 3.1.3). Inspecting the resulting pyrimidine/purine ratio of such an artificially generated sequence, shown in fig. 3.3b, and comparing it with fig. 3.3a, it is obvious that the graphs are very similar to one another. The calculation of α over this sequence (and its different parts, as given below the graphs in fig. 3.3) yields values which are practically equal to those of the λ genomic sequence.

Once this clue was found, investigations in more detail could be done. There-

fore, artificial sequences were generated with a ratio $R(i)$ of pyrimidines/purines remaining constant in intervals of length D, but changing periodically in adjacent intervals, as shown in the graph of fig. 3.4a with D=5000 bp. For the magnitude and changes of $R(i)$ 'realistic' values were chosen being typical for the DNA sequences of the λ-phage and the different myosine heavy chains, as well as many other sequences. To be concrete, in the presented computer simulations, $R(i)$ was chosen to take periodically the values 0.8 and 1.2; cf. [58]. The sequences have been investigated with different interval length D (100 bp-5000 bp) according to the method of the local 'pseudo-fractal' exponent, see section 3.1.2. The length of the parameter L_i remained constant at L_i=5000 bp. The results are summarized in fig. 3.4 and fig. 3.5.

Of particular importance is the exposed dependence of the $\alpha(i)$ graph on the ratio D/L_i. The changes in the overall structure of the corresponding $\alpha(i)$ graphs are striking: One clearly sees how the periodic structure of the $\alpha(i)$ graph disappears and how the high $\alpha(i)$ values continuously decrease with decreasing values of D, or likewise, with increasing frequency of the $R(i)$ variations. In more plain terms, the 'fractal' structure [39] of the sequence (i.e., $\alpha \approx 0.6$) decreases for $D >$500 gradually, and the sequence seems to become completely 'random' (i.e., $\alpha \approx 0.5$) for D=200 or D=100. This interesting quantitative finding may be easily illustrated observing that sufficiently fast fluctuations (or changes) of the pyrimidines/purines ratio $R(i)$ cannot be 'seen' by the calculation procedure of α, since they become 'smeared out'. At this stage the data presented in fig. 3.3 will be reconsidered. According to the present finding, the aforementioned contradictions concerning the 'fractal' or 'random' organization of the DNA sequence appear to be non-substantial.

3.2.4 Dependence of the quantity $\Delta\alpha$ on the coding region

After revealing the weak points of the power law method, it was possible to look for improvements. There were two proposals to overcome the aforementioned problems. One proposed by Peng et al. [66], keeping the interpretation in the conceptual environment of fractals, is called detrended walk, which will be shortly discussed in section 4. The second proposal, made by Dreismann et al. [56], skips all claims on 'fractals' and/or 'long range correlations', and tries to find possibly existing differences of a DNA sequence from randomly generated sequences, using the local 'pseudo-fractal' exponent $\alpha(i)$.

Some first results on this topic can be found in ref. [56], but the detailed quantitative analysis of the fluctuation characteristics of the ratio $R(i)$ and the local 'pseudo-fractal' exponent $\alpha(i)$ along the DNA sequences of living organisms, is presented in ref. [57] and [59].

For illustration of a remarkable difference found, graphs of the adenovirus

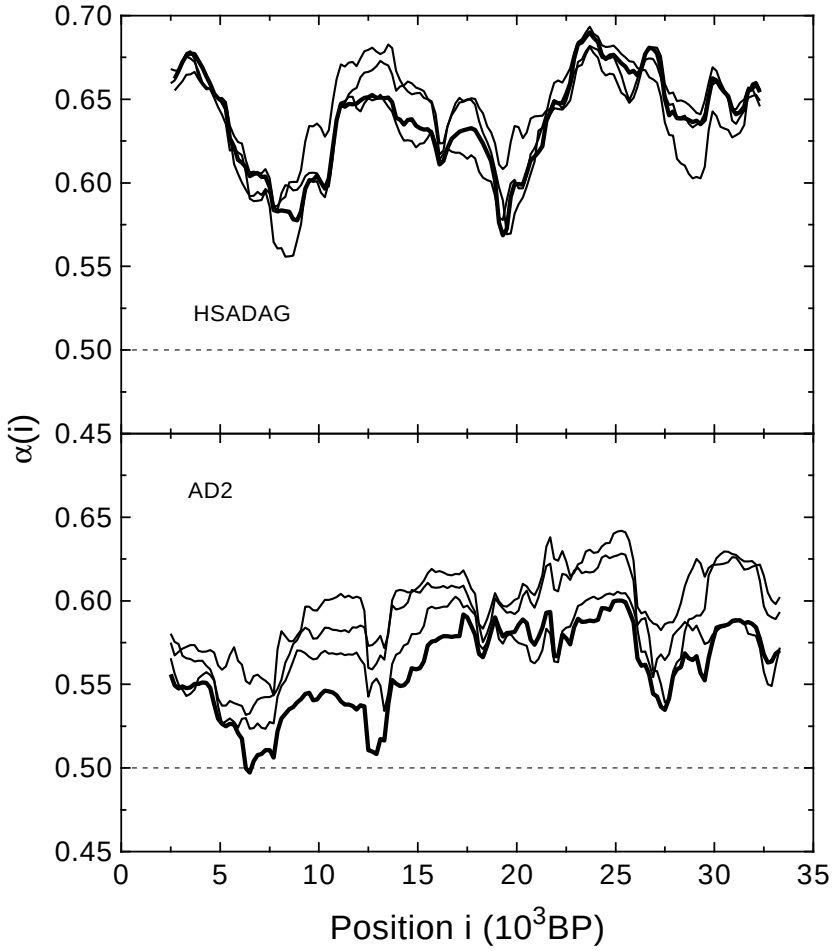

Figure 3.6: The $\alpha(i)$ graphs of the human adenosine deaminase (HSADAG) and the adenovirus type 2 (AD2) (thick lines) DNA sequences and of the associated artificial sequences (thin lines), as generated with our procedure. The parameters are L_i=5000 (L_i : length of subsequence around position i for the calculation of $\alpha(i)$) and k_{max}=500 (k_{max} : maximum correlation length). The pyrimidine/purine ratio $R(i)$ was determined within intervals of 100 bp. The parameter i proceeds along the sequence with steps of size 200 bp.

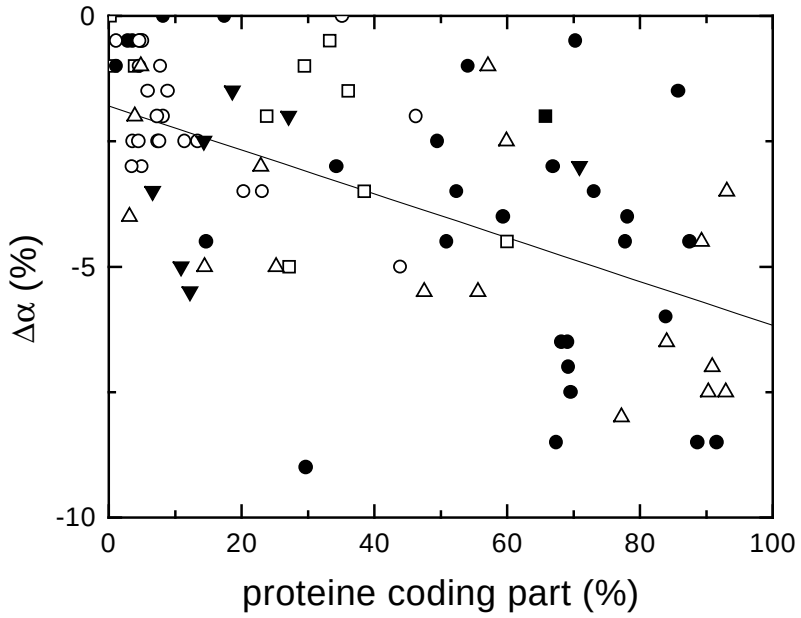

Figure 3.7: Dependence of the quantity $\Delta\alpha$ for nucleotide combination AT/CG on the protein coding region of the DNA sequence. The straight line is fitted according to the standard linear regression method and is only a visual guide. ■ : fungi; • : invertebrates; △ : mammals; + : organelles; × : primates; □ : prokaryotes; ○ : rodents; ♢ : vertebrates; ∗ : phages;

(type 2) DNA sequence and of some typical associated artificial sequences are presented in fig. 3.6. It can be easily recognized that the corresponding $\alpha(i)$ graph of the natural DNA (thick line) takes almost everywhere lower values than those of the artificial sequences (thin lines). This striking finding is the more interesting as it is in clear contrast to the behaviour of the DNA sequences of six human genes where the average value of these differences is zero (cf. table 3.1 and fig. 3.6). The long sequence of yeast chromosome III exhibits still more significant differences of the considered kind from its associated artificial sequences. The average differences are denoted by $\Delta\alpha$ (cf. Eq. 3.6 in section 3.1.4). The DNA sequences under investigation have been quantified as presented in table 3.1. In the corresponding calculations, each natural DNA sequence is compared with ten associated artificial sequences. It is remarkable that all different values of $\Delta\alpha$ vary between 0% and - 10% and it is appropriate to mention that not a single natural DNA sequence has been found so far with a positive value of $\Delta\alpha$.

Trying to clarify the biological reasons and/or origins of this observation, the relation between the differences $\Delta\alpha$ and the percentage of coding parts in the DNA sequences under consideration is investigated. The $\Delta\alpha$ values referring to

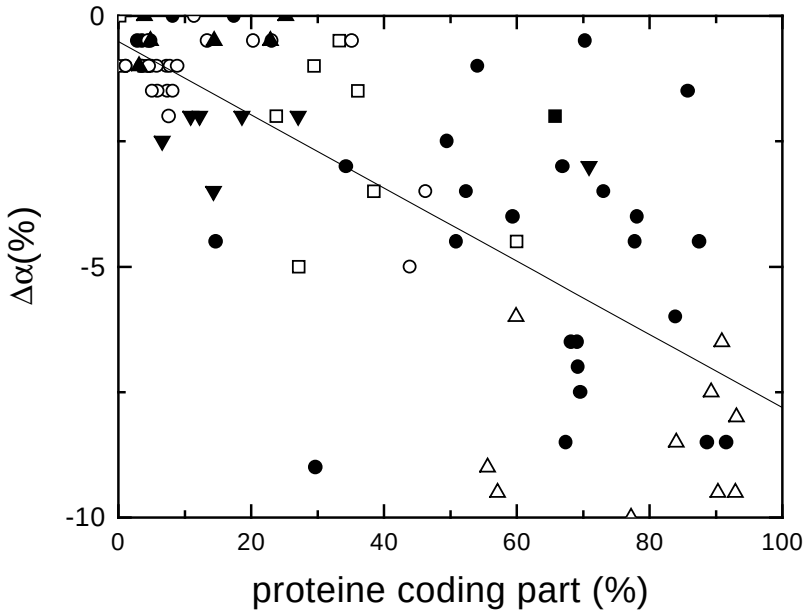

Figure 3.8: Dependence of the quantity $\Delta\alpha$, for nucleotide combination AG/CT, on the protein coding region of the DNA sequence. The straight line is fitted according to the standard linear regression method and should not be misinterpreted as a linear dependence. ■ : fungi; • : invertebrates; △ : mammals; + : organelles; × : primates; □ : prokaryotes; ○ : rodents; ♢ : vertebrates; ∗ : phages;

the nucleotide combination AT/CG are plotted against the percentage of protein coding parts in fig. 3.7, whereas the data points referring to the combination pyrimidine/purine are presented in fig. 3.8 in the same manner. The straight full lines shown are fitted to the data points by the standard linear regression method. It should be emphasized that these lines are guides to the eye. These graphs clearly show that the magnitude (i.e. the absolute value) of $\Delta\alpha$ increases monotonously with increasing percentage of protein coding parts in a DNA sequence. A similar result holds also with respect to the total (i.e., protein, rRNA, etc.) coding parts of the sequences [56, 59].

Nevertheless, the quantities $\alpha(i)$ and α may be simply considered as empirical parameters being associated with, and describing specific characteristics of, the fluctuations of pyrimidine/purine (or other nucleotide combinations) along a DNA sequence, as already the definition of the original function $\alpha^*(k)$ shows. These parameters may be then of certain usefulness for the characterization of the primary structure of DNAs.

Indeed, the following remarks give support to this expectation. The data representing our main result clearly demonstrate the existence of certain DNA se-

quences with specific stochastic characteristics – as quantified by $\alpha(i)$ and $\Delta\alpha$ – being significantly different from those of artificial sequences. To be more specific, the presented results reveal that more negative $\Delta\alpha$ values are related to more compact genomes, which means that the natural sequences exhibit less 'fluctuations' than the associated artificial ones. On the contrary, natural DNA sequences containing a large amount of introns have a very small or a vanishing $\Delta\alpha$, i.e. they appear to be very similar to random generated artificial sequences. In essence, this finding is in contrast to the main point of ref. [39], since e.g. an 'effect', i.e. $\Delta\alpha \neq 0$, is detected for DNAs having a small amount of introns (or non-coding parts), whereas DNAs containing a large amount of introns show no 'effect', i.e. $\Delta\alpha \approx 0$.

This finding is in some sense also biologically interpretable, because it is naturally expected that, if a DNA sequence has certain fluctuation characteristics which carry 'information', then this should happen in coding regions, rather than in introns – and by no means in the opposite way! Of course, this remark does not imply that introns or other non-coding parts do not contain any (still unknown) information at all.

It may also be noted that the negative sign of the $\Delta\alpha$ values should not be considered any more as 'strange', since now it is known that the quantities $\alpha(i)$ represent just empirical parameters, which have less to do with 'self-organization', 'self-similarity', etc.. Crucial, however, is the observation that $\Delta\alpha \neq 0$. Thus far, the possible physical meaning of the negative sign of $\Delta\alpha$ is not clear. One may speculate that there is some possible relation to the existence of isochore regions [71] in DNAs and/or to the actual codon usage of protein coding sequences.

Let us now consider some additional details. In more biological terms, our finding under consideration means that the genomic organization of compact DNAs exhibits less interchanges between the considered two nucleotide combinations (cf. fig. 3.4) than every associated random sequence. The investigated human DNA sequences exhibit values of $\Delta\alpha$ between 0 and -2% in the case AG/CT, and between -1.5 and -3.5% for the nucleotide combination AT/CG. As already mentioned, sofar not a single DNA sequence has been found having a positive $\Delta\alpha$ value. Additionally, inspection of fig. 3.7 gives strong evidence to the fact that the introns in the human DNA sequences under consideration exhibit small but non- vanishing 'fluctuation' characteristics ($\Delta\alpha \approx$ -2%; see the intercept on the ordinate) for the AT/CG combination, whereas, with respect to the AG/CT combination, such a clear-cut result does not exist.

Table 3.1: Summery of all investigated sequences. Given are in column 1) EMBL database identification name (ID); 2) database category of sequence; 3) short description of genome; 4) length of sequence in base pairs; 5) percentage of protein coding region; 6) 'pseudo-fractal' exponent α of the complete sequence for either base combination AT/CG and AG/CT (pyrimidine/purine); 7) value of the measure $\Delta\alpha$ for both base combinations AT/CG and AG/CT.

1) EMBL identification name	2)	3) genome	4) length [bp]	5) [%]	6) α		7) $\Delta\alpha$ [%]	
					AT/CG	AG/CT	AT/CG	AG/CT
SCCEN11D	FUN	S.cerevisiae centromeric region CEN11	24743	66	0.58	0.61	-5	-2
SCCHRIII	FUN	S.cerevisiae chromosome III	315357	67	0.62	0.62	-4.5	-3
CEB0303	INV	Caenorhabditis elegans cosmid B0303	41071	36	0.60	0.61	-2.5	-1.5
CEF59B2	INV	Caenorhabditis elegans cosmid F59B2	43782	-	0.59	0.62	-2.5	-1
CELRP	INV	Caenorhabditis elegans LDL receptor	23719	60	0.58	0.59	-5.5	-4.5
CER08D7	INV	Caenorhabditis elegans cosmid R08D7	27368	33	0.61	0.65	-2	-0.5
CER107	INV	Caenorhabditis elegans cosmid R107	40970	24	0.62	0.63	-3	-2
CEUNC22	INV	Caenorhabditis elegans unc-22 gene	47081	39	0.65	0.64	-6	-3.5
CEZK370	INV	Caenorhabditis elegans cosmid ZK370	37675	30	0.61	0.62	-2	-1
CEZK637	INV	Caenorhabditis elegans cosmid ZK637	40699	-	0.60	0.61	-1	-1.5
CEZK643	INV	Caenorhabditis elegans cosmid ZK643	39534	-	0.64	0.63	-4	-1.5
DMMHC	INV	D. melanogaster myosin heavy chain	22663	27	0.69	0.59	-2.5	-5

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
DMZFP	INV	D. melanogaster zinc finger protein	28627	-	0.71	0.62	-1.5	-2
BTAS1C	MAM	B.taurus alpha-S1-casein gene	22069	-	0.62	0.57	-1.5	-4
BTCASAS2X	MAM	Bovine alpha s2 casein type A protein	21246	-	0.65	0.61	-0.5	-2.5
OCBGLO01	MAM	Rabbit beta-like globin gene	44594	4	0.62	0.66	-2	-1
OCTITINR	MAM	O.cuniculus mRNA for titin	20415	-	0.51	0.47	-15	-12
CHEGZ	ORG	Euglena gracilis Z Chloroplast	41017	44	0.64	0.62	-0.5	-5
CHEVCG	ORG	Epifagus virginiana chloroplast	70028	46	0.71	0.61	-2	-3.5
CHMPXX	ORG	Marchantia polymorpha chloroplast	121024	59	0.72	0.61	-1	-4
CHNTXX	ORG	Tobacco chloroplast	155844	52	0.66	0.60	-4.5	-3.5
CHOSXX	ORG	Rice chloroplast	134525	50	0.66	0.62	-4.5	-2.5
KLKPMAX	ORG	Leishmania tarentolae kinetoplast maxicircle	20991	35	0.67	0.70	± 0	-0.5
KTKPGEN	ORG	Trypanosoma brucei brucei kinetoplast maxicircle	23016	-	0.71	0.73	-2	-1
MIBPCG	ORG	B.physalus mitochondrion	16398	69	0.54	0.59	-7	-6.5
MIBTXX	ORG	bovine mitochondrion	16338	69	0.56	0.60	-4.5	-7
MICCCG	ORG	C.carpio mitochondrion	16364	70	0.53	0.59	-9.5	-7.5
MIHSXX	ORG	H.sapiens mitochondrion	16569	68	0.53	0.60	-5	-6.5
MIMM	ORG	Mouse mitochondrion	16295	30	0.52	0.58	-6.5	-9
MIMPCG	ORG	Marchantia polymorpha mitochondrion	186608	34	0.65	0.60	-1.5	-3
MIPAGEN	ORG	Paramecium aurelia mitochondrion	40469	54	0.67	0.65	-2	-1

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
MIPLCG	ORG	Paracentrotus lividus mitochondrion	15696	73	0.54	0.60	-6	-3.5
MIPVDNA	ORG	P.vitulina mitochondrion DNA	16826	67	0.54	0.60	-6.5	-8.5
MISCCG	ORG	S.cerevisiae mitochondrion	78521	15	0.66	0.59	2	-4.5
MISPCG	ORG	S. pombe mitochondrion	19431	51	0.57	0.60	-4.5	-4.5
MTPACG	ORG	P.anserina mitochondrion	100314	77	0.55	0.54	-8	-10
GCHBEGEB	PRI	Galago crassicaudatus epsilon-, gamma-, delta-, and beta-globin	41101	5	0.63	0.64	-3	-1
GGAFPA	PRI	Gorilla alpha-fetoprotein (AFP)	24607	7	0.61	0.61	-2.5	-1.5
HSADAG	PRI	Human adenosine deaminase	36741	3	0.62	0.66	-3.5	-0.5
HSAFPCP	PRI	Human alpha-fetoprotein	22166	8	0.63	0.59	-2	-1.5
HSATP1A2	PRI	Human Na,K-ATPase subunit alpha 2	26668	11	0.62	0.66	-2.5	± 0
HSBFXIII	PRI	Human factor XIII b	33206	6	0.65	0.64	-1.5	-1.5
HSBMYH7	PRI	Homo sapiens beta-myosin heavy chain	28438	20	0.62	0.67	-3.5	± 0
HSCBMYHC	PRI	Human cardiac beta myosin heavy chain	25000	23	0.62	0.68	-3.5	± 0
HSCC1S14	PRI	Human cosmid clone HDAB	34379	-	0.73	0.63	-1.5	-1.5
HSCGMP4	PRI	H.sapiens cGMP phosphodiesterase	27847	-	0.73	0.74	-1	-1
HSCRYGBC	PRI	Human gamma-B-crystallin and gamma-C-crystallin genes	22775	5	0.64	0.62	-1	-1

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
HSCSF1PO	PRI	Human c-fms proto-onco for CSF-1 receptor	35100	8	0.64	0.65	-3	±0
HSCYP8P	PRI	Human debrisoquine 4-hydroxylase (CYP2D8P) and (CYP2D7) pseudogenes	17060	17	0.70	0.67	-1.5	±0
HSDYSTROP	PRI	Human dystrophin	38770	-	0.60	0.63		-2.5
HSFIXG	PRI	Human factor IX	38059	4	0.62	0.63	-2.5	-1
HSGHCSA	PRI	Human growth hormone	66495	5	0.64	0.64	-2.5	-0.5
HSG6PDGEN	PRI	Human glucose-6-phosphate dehydrogenase	20114	8	0.68	0.63	-2.5	-2
HSHBB	PRI	Human beta globin	73326	4	0.64	0.67	-2	-0.5
HSHDABCD	PRI	Human 3 cosmids (HDAB, HDAC ,HDAD)	58864	-	0.72	0.64	-0.5	-2
HSHDAC	PRI	Human cosmid HDAC	40289	-	0.72	0.67	-1	-2
HSHDAD	PRI	Human cosmid HDAD	40103	-	0.72	0.64	-0.5	-2.5
HSHPRT8A	PRI	Human hypoxanthine phosphoribosyltransferase	56737	1	0.68	0.62	-0.5	-2
HSIFNAR	PRI	Human IFNAR interferon alpha/beta receptor	32906	5	0.68	0.63	-0.5	-1.5
HSIGLAMB	PRI	Human lambda-immunoglobulin	33737	7	0.68	0.64	-2	-1
HSMMDA	PRI	Human cosmid MMDA	37314	-	0.65	0.68	-1	±0
HSMMDBC	PRI	Human cosmid MMDB and MMDC	68505	-	0.67	0.66	-1.5	±0
HSNEUROF	PRI	Human oligodendrocyte myelin glycoprotein	100849	6	0.65	0.67	±0	-1
HSP53G	PRI	Human p53 gene	20303	8	0.63	0.64	-1	-1
HSRCC1	PRI	Human HeLa cell	34641	3	0.65	0.60	-3	-1

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
HSTCRADCV	PRI	Human Tcr-C-delta gene	97634	9	0.63	0.65	-1.5	-1
HSTHB	PRI	Human prothrombin	20801	5	0.66	0.63	-0.5	-1
HSTPA	PRI	Human tissue plasminogen activator (t-PA)	36594	13	0.70	0.63	-2.5	-0.5
HSVWFAA	PRI	Human von Willebrand factor	21352	-	0.70	0.63	-1.5	-2
HSVWFAB	PRI	Human von Willebrand factor pseudogene	21033	4	0.61	0.62	-2.5	-0.5
ARPRITL	PRO	A. rhizogenes Integrated Ri plasmid agropine	21126	57	0.67	0.50	-1	-9.5
ATACH5	PRO	Agrobacterium tumefaciens Ti plasmid pTi15955	24595	48	0.60	0.53	-5.5	-11
AVNIFC	PRO	A.vinelandii major nif gene	28793	90	0.60	0.55	-7.5	-9.5
EC2MIN	PRO	E. coli 2 minute region	28277	56	0.59	0.54	-5.5	-9
ECAPAH01	PRO	E.coli K12 genome	111402	84	0.58	0.52	-6.5	-8.5
ECUW85U	PRO	E. coli	91408	89	0.56	0.50	-4.5	-7.5
KPNIF	PRO	Klebsiella pneumoniaer nif gene	24206	-	0.59	0.50	-5.5	-14.5
MLB1912CS	PRO	M. leprae cosmid b1912	38542	-	0.61	0.48	-5	-11.5
MLB1935CS	PRO	M. leprae cosmid b1935	40123	-	0.55	0.50	-6.5	-12.5
MLB38COS	PRO	M. leprae cosmid b38	37114	-	0.57	0.57	-6.5	-12.5
MLB577COS	PRO	M. leprae cosmid b577	37842	60	0.62	0.53	-2.5	-6
PTREGU	PRO	Atumefaciens Plasmid Ti	29802	93	0.54	0.54	-7.5	-9.5
RCPHSYNG	PRO	R.capsulatus photosynthesis cluster	45959	99	0.45	0.46	-17	-16.5
SEERYAB	PRO	S.erythraea second and third ORF's of eryA	20235	99	0.46	0.46	-16	-15

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
SEERYABS	PRO	S.erythraea eryA gene	20444	93	0.68	0.52	-3.5	-8
STRFB	PRO	S.enterica rfb gene	22080	91	0.57	0.55	-7	-6.5
VCRFBAT	PRO	V.cholerae genes for rfbA-rfbT, ompX, ?orf1-3	20134	-	0.58	0.60	-8	-10
MMBGCXD	ROD	Mouse beta-globin	55856	14	0.58	0.61	-5	± 0
MMPACOLL	ROD	Mouse pro-alpha1 (II) collagen chain	30701	25	0.60	0.68	-5	-0.5
MMSLPSEXB	ROD	Mouse nonfunctional sex-limited protein	26307	3	0.61	0.63	-4	-0.5
MMTCRA	ROD	Mouse T-cell receptor	94647	-	0.63	0.64	-1.5	± 0
MMTCRVAD	ROD	Mouse T-cell receptor (TCR V-alpha 161)	34476	5	0.77	0.76	-1	-0.5
RNCRYG	ROD	Rat gamma-crystallin gene	54670	-	0.71	0.63	-1.5	± 0
RNIGF2	ROD	Rat IGFII insulin-like growth factor II	30000	23	0.60	0.63	-3	1
RNSCIII	ROD	Rat brain sodium channel III	6822	86	0.56	0.61	-5.5	-1.5
RNSCPIIR	ROD	Rat brain sodium channel protein II	8553	70	0.62	0.62	-3	-0.5
ADRCOMPGE	VRL	Mastadenovirus h5	35935	71	0.69	0.64	-3	-3
AD2	VRL	Adenovirus type 2	35937	78	0.68	0.59	-2.5	-4.5
EBV	VRL	Epstein-Barr virus (EBV)	172281	12	0.58	0.58	-5.5	-2
HECHCCOMG	VRL	Channel catfish virus	134226	79	0.68	0.61	-3	-5
HEHCMVCG	VRL	Human Cytomegalovirus Strain AD169	229354	87	0.69	0.58	-3	-4.5
HEHCMVU	VRL	Human cytomegalovirus (HCMV)	43275	89	0.60	0.58	-3.5	-5.5
HEHS1ULR	VRL	Herpes simplex virus type 1	108360	69	0.70	0.68	-3.5	-2.5
HEHS1US	VRL	HSV1 (strain 17)	26245	-	0.70	0.64	-3	-3.5

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
HEHS4B958	VRL	Epstein-Barr virus, artifactual joining of B95-8	184113	32	0.61	0.57	-3	-3.5
HEHS5VF	VRL	Human cytomegalovirus F fragment	20349	96	0.53	0.52	-4	-8.5
HEHS6U111	VRL	Human herpesvirus 6 major capsid protein	24927	83	0.67	0.57	-7	-6.5
HEHSECOMG	VRL	Equine herpesvirus 1	150223	89	0.57	0.60	-3.5	-6.5
HEHSV3PRG	VRL	Herpesvirus saimiri	43658	87	0.58	0.62	-7	-6.5
HEVZVXX	VRL	Varicella-Zoster virus	124884	89	0.65	0.54	-4.5	-8.5
HE1CG	VRL	Herpes simplex virus (HSV) type 1	152260	78	0.64	0.63	-3	-4
HSGEND	VRL	Herpesvirus saimiri	112930	94	0.54	0.53	-7	-10
IBACGB	VRL	Avian infectious bronchitis virus pol protein	27608	-	0.52	0.52	-14	-10
IBAORFAB	VRL	Avian infectious bronchitis virus F1 and F2 genes	20500	99	0.56	0.52	-15	-11.5
MUCGENE1	VRL	Murine coronavirus ORF 1a	21798	87	0.55	0.59	-10	-7
PXVACCG	VRL	Vaccinia virus,	191737	-	0.56	0.55	-7.5	-6.5
LAMBDA	PHG	bacteriophage lambda	48502	84	0.69	0.61	-6	-6
PODOT7	PHG	bacteriophage T7	39936	92	0.52	0.52	-10	-8.5
GDCOL6A2G	VRT	Chicken Col6A2 gene	27443	12	0.58	0.58	-5.5	-2
GDLIPLIP	VRT	Chicken lipoprotein lipase	22257	7	0.67	0.58	-3.5	-2.5
GGCRYDS	VRT	Chicken delta-1 and delta-2 crystallin genes,	25342	11	0.62	0.58	-5	-2
GGMYHE	VRT	Chicken embryonic myosin heavy chain gene	31111	19	0.67	0.65	-1.5	-2

continued on next page

continued form previous page

1)	2)	3)	4)	5)	6)		7)	
GGVITIIG	VRT	Chicken vitellogenin II gene	20343	27	0.63	0.59	-2	-2

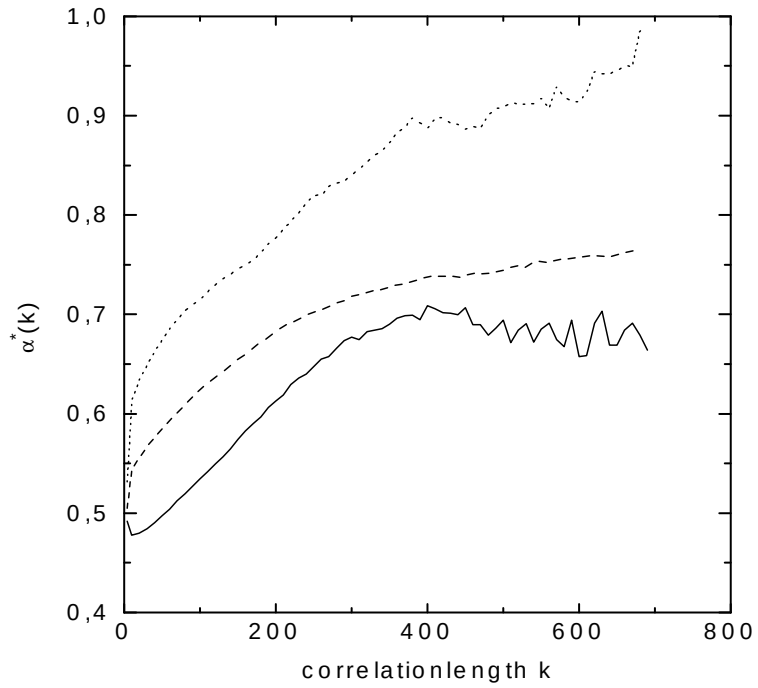

Figure 4.1: $\alpha^*(k)$ graphs using the 'detrended walk'. All graphs are curved and monotonously growing. Consequently the fluctuation $F(k)$ of the 'detrended walk' is not linear in a double logarithmic plot. Presented are $\alpha^*(k)$ of LAMBDA (solid line), SCCHRIII (dashed line) and HSHBB (dotted line). Acronyms are EMBL database identification names (cf. table 3.1).

4 On the 'Detrended Fluctuation Analysis'

In this section the results of the newly presented 'detrended fluctuation analysis' or DFA [66, 67], which – according to the authors of ref. [67] – is an improved method for the detection of long-range correlations in patchy DNA sequences, is shortly discussed. This method intends to take into account the aforementioned changes of base-composition along a natural DNA sequence, thus avoiding the artifacts being caused by the nucleotide patchiness.

The main aspect of the DFA may be described as follows: First the entire sequence of length N is divided into N/l non-overlapping boxes. Each box contains the number of l nucleotides. Secondly, a 'local trend' or slope in each box is determined by a linear least-square fit for the DNA walk displacement $y(i)$, as mentioned in section 3. Next, the 'detrended walk' displacement $y_l(i)$ is defined in that box as the difference between the original walk $y(i)$ and the local trend or

slope. Then, the variance of this 'detrended walk' is calculated, and the definition of a 'scaling exponent' in analogous way as in the case of the random walk of section 3; see [66, 67] for details.

The main idea behind this procedure (which is clearly more involved than the original one [39]) is given by the mentioned difference between the original walk $y(i)$ and the local trend. With this method, and choosing an appropriate box length l, one hopes to 'remove' the natural variations of base-composition along a DNA sequence.

After the numerical implementation of the DFA method, the associated 'detrended' exponent $\alpha^*(k)$ for a large number of DNA sequences is calculated. The so far calculated graphs $\alpha^*(k)$ appear to be rather 'curved' functions of the correlation length k; see fig. 4.1. Therefore it may be concluded that, also in this case, there are no well-defined 'scaling' or 'fractal' exponents.

5 On 'Linguistic' Tests

5.1 Theoretical remarks

In order to apply the 'linguistic' tests [60], mentioned in the introduction, to DNA sequences, the concept of 'word' has to be introduced. Of course, in the case of coding sequences, the biologically relevant 'words' are the well-known triplets, which code for amino acids according to the (almost) universal genetic code. For non-coding regions of DNA, however, biologically relevant 'words' are not known, if they exist at all.

Therefore, Mantegna et al. [60] considered n-tuples, where n is a free parameter between 3 and 8. To obtain the different n-tuples needed to perform the 'linguistic' analyses, a 'reading window' of length n is shifted progressively along the DNA sequence of interest at one base. Note that there are 4^n different n-tuples, since there are 4 'letters' (i.e., A,G,C,T) in the 'alphabet' used by DNA.

To implement the 'linguistic' test as given by the Zipf analysis [61], all the 'words' are ranked (in the present case: of a given length, i.e. the n-tuples) in the order of their actual frequency of occurrence in a given DNA sequence. It is then convenient to make a double-logarithmic histogram, plotting the logarithm of the frequency of occurrence of an n- tuple against the logarithm of its rank. According to the original claims [60], it then surprisingly appears that the produced graph is linear over a significant range of the rank. (E.g., if n=6, the linearity extends from rank 1 to roughly rank 1000). The used 'word' lengths were between 3 and 8. This linearity is considered to be the characteristic feature of the so-called Zipf's law [61]. The slope to the graph (if it is linear) is called *Zipf exponent*.

The same n-tuples are also needed for the second 'linguistic' test of ref. [61], which is based on Shannon's information-theoretical concept of entropy [62, 63]. According to Shannon, the entity *'information'* is directly associated with 'reduction of entropy'. Related to this reduction is also another quantity of information theory, called *redundancy*. In simple terms, redundancy is the degree to which a given text, which represents an 'information', can be understood even when letters are missing and/or incorrect. Therefore this quantity is also a measure of the flexibility of a 'language' or a 'code'.

The mathematically precise definitions of these quantities are as follows [60, 62, 63]. The entropy (or better: the n-entropy) $H(n)$ is given by

$$H(n) = -\sum_{i=1}^{4^n} p_i log_2 p_i \tag{5.1}$$

where n is the (constant) length of all 'words'. The redundancy Re is defined

through a limes, i.e.

$$Re = \lim_{n\to\infty} Re(n) \tag{5.2}$$

with

$$Re(n) = 1 - \frac{H(n)}{kn} \tag{5.3}$$

where, by convention, $k = log_2 4 = 2$ (see e.g. [60]). The maximum value of n for which it is possible to determine the n-entropy appears to be $n = 6$. For larger n-values, many possible words are rarely present, i.e. they exhibit extremely bad statistics which obscure the numerical values of $H(n)$ and $Re(n)$.

As mentioned above, it was claimed [60] that these two 'linguistic' tests reveal significant differences between coding and non-coding parts of natural DNA sequences. Furthermore, is was found that the analyzed non coding DNA sequences exhibit larger values of redundancy than did the coding DNA sequences, which suggests – as Mantegna et al. in [60] put it – "the possible existence of one (or more than one) structured biological language(s) present in non-coding DNA sequences".

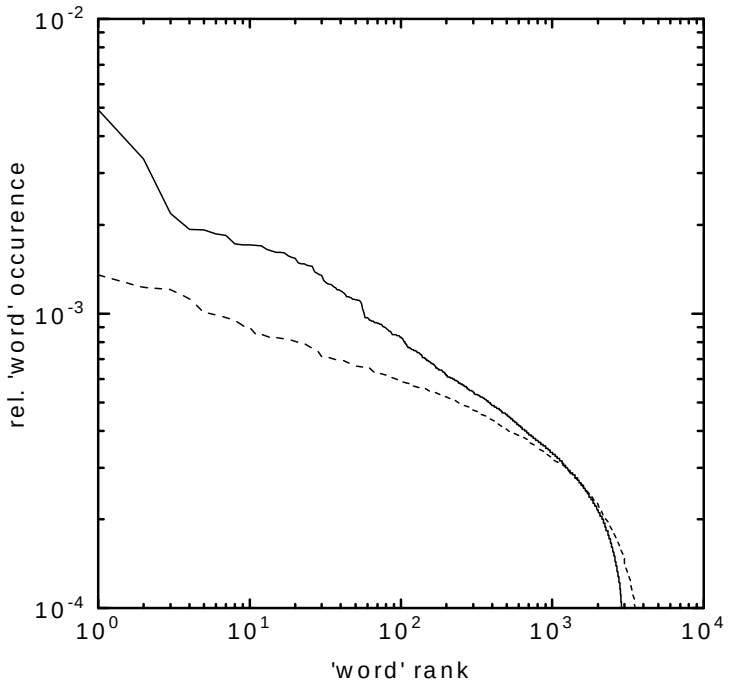

Figure 5.1: 6-tuple Zipf plot of the human sequence HSRETBLAS (solid line) and E.Coli sequence ECUW89 (dashed line), which are studied in [60]. It is difficult to decide on linearity in double logarithmic plots. For a linear presentation see fig. 5.2. Acronyms are EMBL database identification names (cf. table 3.1 on pages 28ff).

5.2 Results

5.2.1 Linearity of Zipf graphs

To check the claim concerning the linearity of the Zipf graphs, i.e. the Zipf-like scaling behaviour, the Zipf plots of different coding and non-coding DNA sequences are calculated and displayed graphically (cf. fig. 5.1). The main features of these graphs are qualitatively in agreement with those displayed in the figs. 1 and 2 of ref. [60]. To be concrete, here are (in fig. 5.1) the corresponding graphs, for 6-tuples, of the human sequence HSRETBLAS (1.5% coding) and the E. coli sequence ECUW89 (82.1% coding). The DNA sequences presented are also studied in ref. [60]. However, it should be recognized that these plots are double-logarithmic, which makes it very difficult to assess quantitatively whether the slope is really constant or not. Therefore, the slopes of these graphs have also been calculated numerically, which are now displayed in a linear scale in fig.

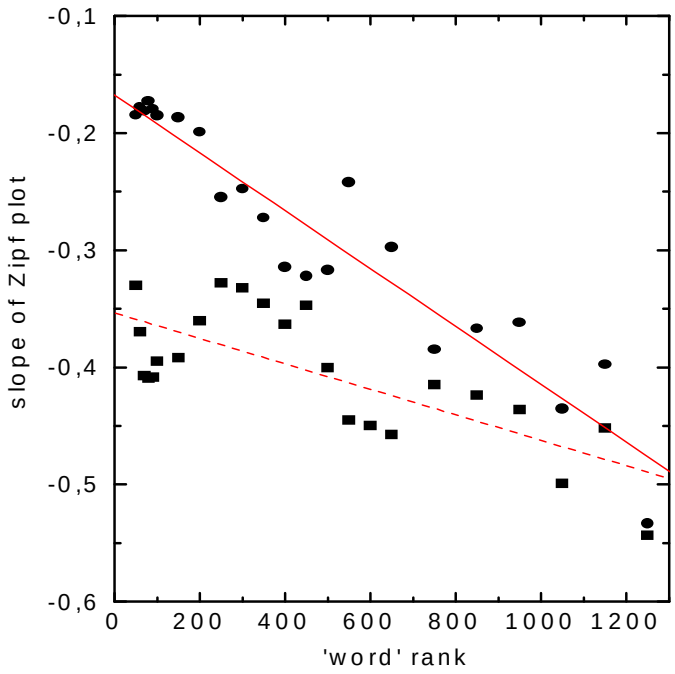

Figure 5.2: Local slope of the Zipf plots in fig. 5.1 proving them to be curved. The slope has the tendency to decrease with increasing 'word' rank and therefore the graphs do not fit Zipfs law as claimed [60]. Every data point represents the slope of 100 'words' in the Zipf plot, which are fitted according to the standard linear regression method. The lines are only guides to visualize the tendency in slope. ■, dashed line: ; ●, solid line.

5.2, together with the corresponding Zipf plots (fig. 5.1). It is obvious that these slopes are not constant; they clearly exhibit a curved, monotonously increasing behaviour. Similar results are obtained for almost all DNA sequences analyzed. In summary, the present investigation fails to find constant slopes of the claimed extension, i.e. over about three orders of magnitude.

5.2.2 Zipf graphs of coding and non-coding DNA sequences

The DNA sequences studied here show the following qualitative difference: The graphs of the non-coding sequences are usually 'steeper' than those of the (mostly) coding ones. This result supports qualitatively the finding [60] that the Zipf slope (or: the Zipf exponent) is larger – by about 50% – for the non-coding sequences. But an exception is also found, as the Zipf graph of the herpesvirus saimiri DNA

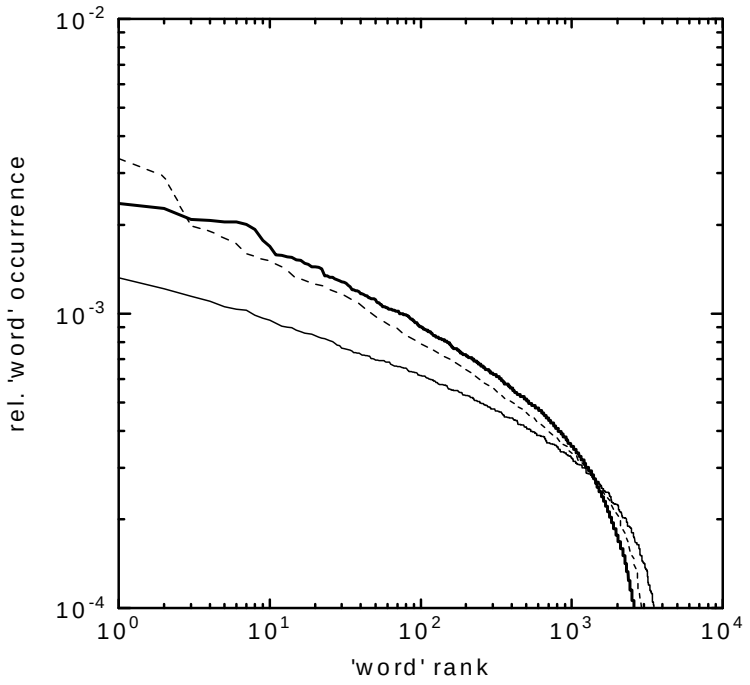

Figure 5.3: Averaged Zipf graphs of 5 mostly coding (thin solid line) and 5 nearly non-coding (thin dashed line) sequences. In addition the Zipf plot of the herpes virus saimiri (HSGEND) is presented, which is highly coding (94%). The 'word length' is 6 bp. The figure reveals a difference in averaged Zipf plots of coding and non-coding sequences, but disproves the difference to be general for individual DNAs.

sequence (HSGEND, 94% coding), shown in fig. 5.3, clearly demonstrates. The overall behaviour of this graph is identical to that of a typical (mostly) non-coding sequence. This proves that the claimed difference [60] between coding and non-coding DNA sequences is not universal.

5.2.3 Zipf tests of natural DNAs and computer simulated sequences

Nevertheless, in most cases studied, there are indeed visible differences between the averaged slopes of the Zipf graphs: coding graphs appear to exhibit a smaller slope – on the average – than non-coding DNA's, in agreement with ref. [60]; for some examples, see fig. 5.4. This qualitative finding motivates the questions whether the observed differences have a biological significance, and how they can be quantified properly.

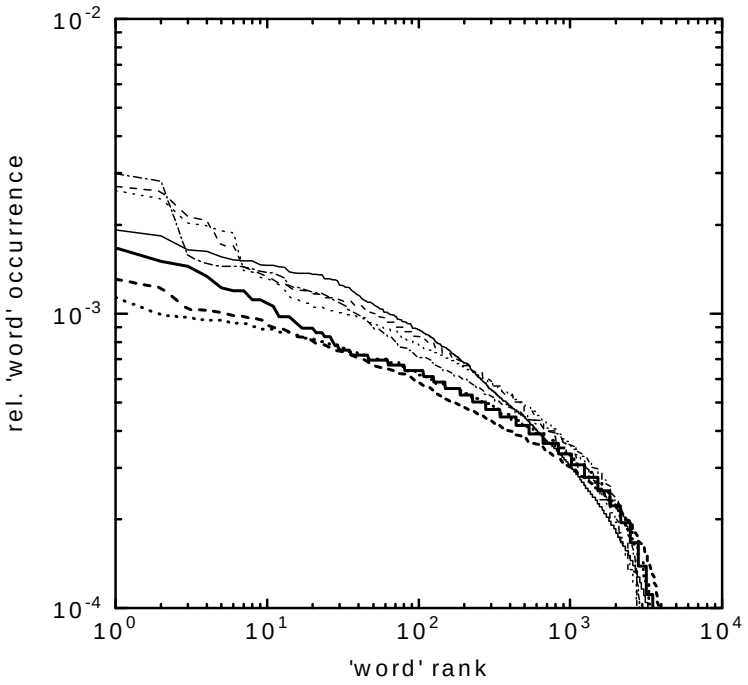

Figure 5.4: 6-tuple Zipf plots of different highly coding (thick lines) and mostly non-coding (thin lines) sequences. Shown are EBV (solid thin, 12%), RNIGF2 (dashed thin, 23%), HSGHCSA (dotted thin, 5%), HSHBB (dashed-dotted thin, 4%), AD2 (solid thick, 78%), HEVZVXX (dashed thick, 89%) and LAMBDA (dotted thick, 84%). The given acronyms are the EMBL-database identification names (see table 3.1). In parentheses first the line shape in the figure and second the coding percentage of the sequence.

Trying to clarify these questions concretely and *quantitatively*, the same 'linguistic' Zipf analysis is applied on a large number of artificial, computer-generated sequences using random number generators; for details see refs. [56, 57] and section 3 above.

First of all, artificial sequences were produced with constant bp composition along each sequence, but of different lengths. The analysis of these sequences shows directly that the Zipf graphs of the shorter sequences have a larger slope than that of the longer sequences. Furthermore, the overall shape of these graphs is clearly reminiscent of the shapes of Zipf graphs of natural DNAs, although the slopes of the artificial sequences seem to be smaller than those of the natural DNAs; see fig. 5.7 and compare with figs. 5.3, 5.4.

In the light of this qualitative result, it may be wondered about the 'reasons'

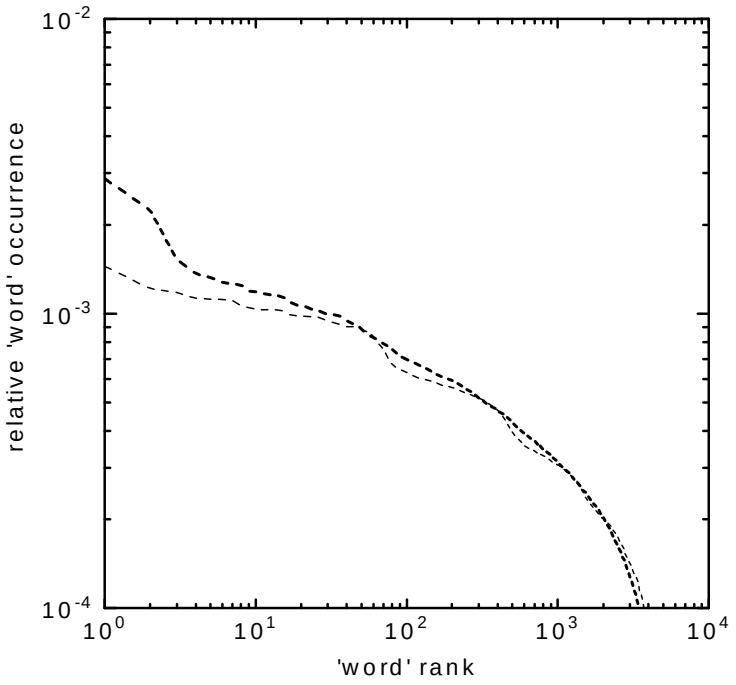

Figure 5.5: 6-tuple Zipf plot simulation of the yeast chromosome III (SCHRIII). The Zipf plot of the associated artificial sequence (thin line, $D : i$=200; cf. section 3) is nearly indistinguishable from the original sequence(thick line).

and/or 'origins' of the observed specific form of the Zipf graphs discussed above: Can some 'biological information' be associated with the observed forms of the Zipf graphs, as suggested in [60]? Or are these graphs related to some, thus far unknown, 'numerical artifacts'?

In order to clarify these questions, the same Zipf analysis has been applied to a large number of computer generated artificial sequences – being associated with a given natural DNA – of the following specific kind: Every produced artificial sequence has the same length and the same base pair (bp) composition as the associated natural DNA. Moreover, in every chosen interval D_i (with a typical length of, say 100 bp; see below) around any base position i, both natural and artificial DNAs have almost the same bp composition- i.e. the same composition, up to the natural statistical deviations caused by the finite length of the chosen interval D_i.

The most surprising feature of the computer-simulation results is demonstrated in figs. 5.5, 5.6. In fig. 5.5, the Zipf analysis of the complete yeast chromosome III sequence is presented. It can immediately be seen that the Zipf analyses of both

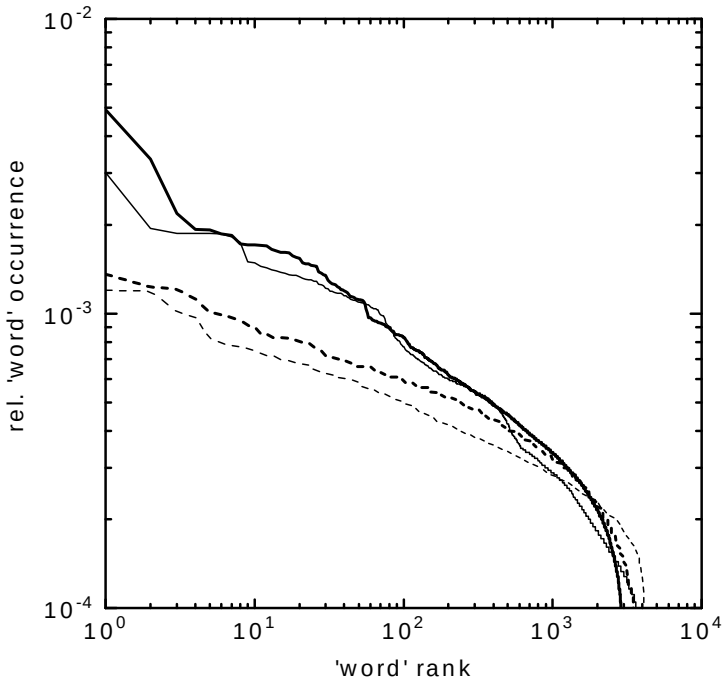

Figure 5.6: Simulation of 6-tuple Zipf graphs of natural DNAs (thick lines) and associated artificial sequences (thin lines). The artificial sequences are generated (cf. section 3) with the parameter D_i as follows: ECUW89 D_i=20 bp (dashed lines) and HSRETBLAS D_i=100 bp (solid Lines). Interesting is the difference of the parameter D_i and the quality of reproduction. For the explanation see the text.

natural and associated artificial sequences (for, say, D_i =200 bp), using 6-tuples as 'words', produced essentially indistinguishable graphs. Essentially the same 'negative' result was obtained for many different natural DNAs, among others: (1) the human HSRETBLAS (cf. above), and (2) the E. coli sequence ECUW89 (cf. above); see fig. 5.6.

These results demonstrate quantitatively that the Zipf analysis [60] is not able to discriminate natural DNAs from the associated computer generated sequences, which furthermore strongly indicates that the Zipf analysis is not able - or not appropriate - to reveal any new biological information being coded in DNA.

5.2.4 Dependence on base composition and patchiness

Although the Zipf graph of every natural DNA can be sufficiently well approximated with the Zipf graphs of associated artificial sequences, as described above,

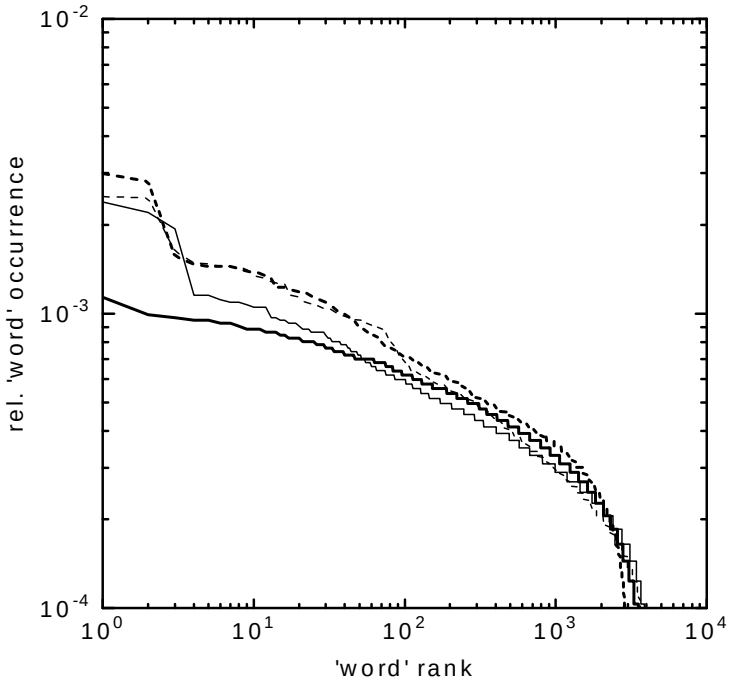

Figure 5.7: Difference in quality of Zipf plot simulation in dependence of the simulation parameter D_i. Presented are the Zipf plots of the human sequence HSHBB (D_i=100 bp; dashed) with a biased base composition and of the λ-phage genome (D_i=15 bp: solid) with an unbiased base composition. To achieve a good simulation a much shorter D_i has to be chosen for a sequence with unbiased base composition than for a sequence with a biased one. Graphs of natural DNAs are thick, those of artificial sequences thin.

it could be that non-coding and coding DNAs exhibit quantitative differences in the quality of this approximation. Namely, it is easily recognized that some DNAs can be approximated (in the considered manner) using larger D_i values than in other cases. See e.g. fig. 5.7, where it is shown that for the 'approximation' of the λ-phage (84% coding) a much smaller D_i value is suitable than in the case of the human β-globin DNA sequence HSHBB (4% coding).

However, further analyses revealed the following unexpected feature: It is not the coding or non-coding character of a natural DNA, which is directly related to the quality of its approximation with artificial sequences, but simply its base composition! Namely, all DNAs studied thus far revealed that natural DNAs which have unequal (or biased) mean base composition (i.e., the frequencies of the occurrences of A, G, C and T in the DNA are not about 25% each) can be

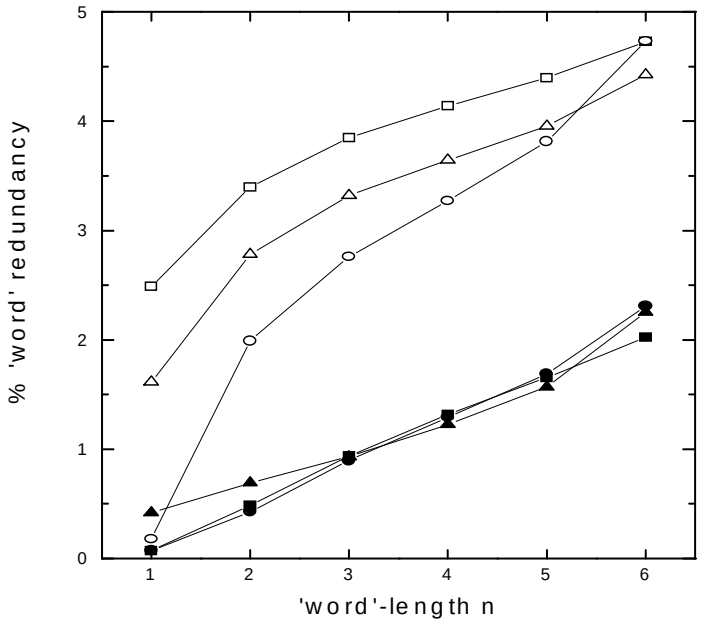

Figure 5.8: Shannon 'word' redundancy test for highly coding (full marks) and mostly non-coding (open marks) sequences. It is remarkable to find no specifics about 'word'-length three even in highly coding sequences, which could be expected, due to the universal genetic code. ■ : ECUW89; ● : LAMBDA; ♢ : AD2 ○ : HSADAG; □ : HSRETBLAS; △ : HSHBB; Acronyms are EMBL database identification names (cf table 3.1).

approximated with artificial sequences, choosing relatively large D_i values (say, some hundreds). On the contrary, if each base has relative occurrence of about 25% in a natural DNA (which may be called an 'equal base composition'), much smaller D_i values (say, some tens) have to be chosen until the aforementioned 'approximation' becomes satisfying.

This surprising finding, in addition to the result presented under 5.2.3 above, indicates that the 'differences' between the Zipf graphs of different natural DNAs may have no biological significance.

5.2.5 Shannon redundancy analysis

The numerical investigations based on Shannon's redundancy $Re(n)$ concept (see above) produced graphs being similar to those presented in ref. [60]. For an

example see fig. 5.8.

The following observation is – from the biological viewpoint – crucial: The triplets are already known to be the 'relevant' words in coding DNA sequences, i.e. they have a well established biological meaning related to amino acid coding. Therefore, a successful 'linguistic' test clearly has to show the specific character of 3-tuples, as compared with 2-tuples, 4-tuples etc. An inspection of the redundancy graphs presented in ref. [60] and fig. 5.8, however, does not satisfy this demand. To be more specific, all $Re(n)$ graphs are just smooth monotonous functions of the word length n, and they exhibit no specific feature at $n = 3$.

Based on this consideration, it may be concluded, in contrast to the claims of ref. [60], that the currently considered quantity $Re(n)$ is not appropriate to reveal any new biological information in non-coding DNA sequences.

6 Biological Origins

To search for possible *biological* reasons for the aforementioned findings, like 'long-range correlations' and 'linguistic structure', the biological features of all sequences given in table 3.1 (pages 28ff) are investigated in detail, and also those of the human β-cardiac myosine heavy chain (MHC) cDNA [58], for which pseudo-fractal α-values differing substantially from those of ref. [39] are obtained.

As discussed in section 3 above, the pyrimidine/purine ratio $R(i)$ (at base positions i along a DNA sequence) plays a crucial role. To calculate $R(i)$ simply an interval (or width, D_i) is chosen around the i-th position and to count the corresponding bases therein. In the case of the bacteriophage λ genome, which is discussed in detail, three distinct regions appear in which $R(i)$ fluctuates around different nearly constant values. Calculation of the mean value α of the pseudo-fractal exponent $\alpha^*(k)$ within each of these regions yielded values which are significantly lower than the mean value α = 0.61 of the complete sequence. However, for the regions of the λ genome that contain the points where the ratio changes, α took higher values; see fig. 3.3. The same features show also the human cardiac β-MHC cDNA sequence, which has two distinct regions with quasi- constant $R(i)$ values.

These as well as additional sequence analyses demonstrate that the value of α correlates with variations of $R(i)$ in the sense that sequence regions with quasi-constant $R(i)$ (homogeneous pyrimidine/purine composition) give $\alpha \approx 0.5$, whereas sequence regions with varying $R(i)$ give higher values of α. Karlin and Brendel [53] have presented a mathematical analysis of the connection between patchiness - which is similar to our $R(i)$ variations - and long-range correlations, and showed, in agreement with our work, that the latter are a trivial consequence of the former. Peng et al. [39] reported that analysis of some intron-containing sequences using regions between minima and maxima ('min-max partitioning') did not change the exponent α. This is just due to complex $R(i)$ variations in these segments, producing numerous local minima and maxima.

All these findings, however, left open the question about the (possible) *biological* origins of the correlations. The first suggestion was that long-range correlations are due to repetitive patterns in introns [43, 44, 50]. Our analyses of the aforementioned as well as of artificial sequences (see Sections 3 and [56, 57]) demonstrate that repeats per se are not necessary to give high pseudo-fractal α values.

Additionally, the almost 50 tandem copies of a 17-mer repeat in intron 2 of the coagulation factor VII gene (studied by Li in ref. [43]) are not by themselves capable of generating a high α value. In certain cases, however, repeats will be able to generate α values which deviate from 0.5. For instance, pyrimidine-rich repeats like poly(C) or poly (CT) will generate such α values if the repeat

units appear at sufficiently large distances from each other, or if they alternate with purine-rich repeats [56]. It is well known that introns and intergenic sequences display numerous tracts of simple nucleotide repeats of varying length, e.g. $(\mathrm{AGGGCTGGAGG})_n$ [72]. The β-cardiac myosine heavy chain gene, for instance, contains in its introns several simple repeats which differ from each other in their pyrimidine/purine ratio.

Thus, the finding called with mathematical terminology 'long-range correlations' may appear because a pyrimidine-rich region is present in one place at some distance from a purine-rich region. However, this does not mean (see e.g. [53]) that the regions influence each other somehow, although it has been interpreted to mean just this ([46], see also [48]). Furthermore, the finding that simple repeats may account for the long-range correlations would still leave open the important question regarding which biological function the simple repeats and/or the long-range correlations might have.

It should be stressed that the simple repeats are highly variable and may differ in length (i.e., in the number of repeat units) even between individuals within the same population [72]. Indeed, such repeats seem to be the most variable parts of genomes and may in fact be used to identify individual members of a population because each individual will display a unique pattern of simple repeats (so called *mini*-satellites) [72]. The variability is probably due to errors introduced by DNA polymerases during DNA replication [73]. The great variability makes it unlikely that most simple repeats that are dispersed in genomes have a well defined biological function. Repeats at centromeres and telomeres, in contrast, do have functions, but such repeats are not present in any of the sequences analyses here or in the literature cited.

A more complex type of dispersed repeats listed as a potential cause of long-range correlations, the primate Alu repeats (cf. ref. [53]), quite clearly has no biological function but can be regarded as selfish or parasitic DNA [74]. Such small interspersed nucleotide elements (SINEs) differ extensively between quite closely related organisms such as the different mammalian orders. Thus, it appears as if the simple repeats and SINEs do not have a function of their own, and their variability, both in structure and evolutionary distribution, further suggests that they are not present for the purpose of providing any long-range correlations. This leaves the long-range correlations as a passive reflection of mostly non-functional DNA segments.

As mentioned above, this investigation among others has noted certain types of long-range correlations also in intron-less sequences; see e.g. [51–53]. The present work shows that in all cases the correlations are related to changes in the ratio $R(i)$ along the DNA sequences. However, changes of $R(i)$ along intron-less DNA sequences may be less conspicuous and are likely to have different biological causes than simple repeats. Here some examples are considered, which

have also been studied by Peng et al. [39].

The case of the phage λ genome has already been discussed above and in section 3. The mentioned three regions with different pyrimidine/purine ratio $R(i)$ correlate with the direction of transcription [70]. Thus, the coding strand has a higher purine content throughout this genome. To test the consequence of transcriptional direction, an inversion of the central region has been performed by a computer, so that the entire genome would be transcribed in the same direction. This produced indeed a mean value β = 0.53. This preponderance of purine on the coding strand of the λ genome presumably relates to codon preferences.

The cardiac β myosine heavy chain cDNA sequences [39] show a change in the ratio $R(i)$ around position 2,500 bp [58]. This change correlates precisely with the boundary between the two domains of the MHC protein. The MHC is composed of two distinct domains with characteristic features [75]. The amino-terminal 850 amino acids comprise a globular domain with a high content of proline and the aromatic amino acids tyrosine and phenylalanine, all of which have pyrimidine-rich codons. The remainder of the protein constitutes an α-helical rod practically devoid of proline and low in phenylalanine and tyrosine. The rod domain has a preponderance of glutamate which utilizes codons with purines only. Thus, the different pyrimidine/purine ratios of the two parts of the MHC cDNA sequence are a consequence of evolutionary selection for the amino acid composition of the corresponding protein domains. For the four- fold degenerate amino acids, the proportion of codons with pyrimidines in the wobble position is approximately the same between the two domains. This shows that choice of wobble base does not contribute to the difference in base composition, strongly indicating that codon preferences rather than some inherent property of the DNA sequence determine pyrimidine/purine ratio along the DNA strand. All MHC sequences, regardless of animal species, display these characteristic features [75]. The reason why the embryonic MHC sequence [39] gives a different (and smaller) pseudo-fractal α value than all the other MHC sequences (which seem to be 'fractal'), is because it is incomplete and only contains the part encoding the rod domain; thus it does not contain the aforementioned point of change of the ratio $R(i)$ [58].

Also the cDNA sequences of chicken c-myb and human dystrophin contain regions with different nucleotide ratios which correlate with the amino acid composition of the different parts of the proteins. Other intron-less sequences are more homogeneous in their nucleotide composition as exemplified by a large number of yeast intron-less genes described in ref. [51].

The human mitochondrial genome [58], finally, is composed of multiple intron-less genes. Here, the ribosomal genes display a value of $R(i)$ which is quite different from the genes that encode proteins (results not shown). Also, the postulated URF6 gene differs slightly in $R(i)$ from its flanking genes NDS and cytochrome b. The URF6 gene is on the opposite strand from the other two [58].

These findings suggest that it is not necessary to assume the pyrimidine/purine dichotomy in order to get α-values larger than 0.5. To test this, the same quantity $F(k)$ has been computed by considering (G,C) versus (A,T), as G and C form a Watson-Crick base pair with a different number of H-bonds than the A-T base pair. Indeed, completely analogous results with those presented above are obtained for pyrimidines versus purines [58].

The possible biological significance of the Voss study [45] cannot be assessed, since its results are averages over complete GenBank database categories, thus discarding the fact that these categories are not of equal taxonomic rank. Moreover, our current (preliminary) investigations show that the β-values [45] of individual sequences within one evolutionary (or database) category differ much more than the average β- values of different categories, thus making the use of β for characterization of *individual* sequences useless.

Concluding, it may be noted that the above considerations strongly indicate that the long-range correlations under consideration are simply reflections of base-composition fluctuations in natural DNA.

7 An Application of the Information Theoretical Kullback Measure on DNA

7.1 Theoretical remarks

In statistics, the problem of reoccurrence of an arbitrary event is well known, either for events depending on a continuous parameter (as in the case of the well known Poisson distribution) or for 'discrete' events (like the occurrence of one of the bases A, C, G, T in a DNA sequence). E.g. in a random sequence of bases, the statistics for the latter case is described by the negative binomial distribution:

$$w_{theo}^{dist} = p(1-p)^{dist} \tag{7.1}$$

p represents the probability of a single event, e.g. the occurrence of a specific base in this random sequence, and w_{theo}^{dist} the associated theoretical distribution of distances $dist$ between two consecutive occurrences of the base under consideration. For the following investigations, it is necessary to evaluate the 'experimental', i.e. the actually appearing, distribution w_{exp}^{dist} of distances $dist$ (where $dist = 0, 1, 2, ...$) in a natural DNA sequence. This is accomplished by counting the number of other bases between two occurrences of the specific base of interest. E.g., in 'ATCCTGGA' the distance between the two adenines is $dist = 6$ and between the thymines is $dist = 2$. If there are adjacent occurrences of the base of interest, like the two guanines in the example, $dist$ equals zero ($dist = 0$).

Finally the 'experimental' and 'theoretical' distribution of distances are compared. Approaches to this problem are often related to the famous *'Shannon entropy'*. In these investigations, the *Kullback* measure [59, 68] is explicitly used in order to achieve this comparison:

$$Ku(w_{theo}, w_{exp}) = \sum_{i=0}^{maxdist} w_{exp}^{i} \ln \frac{w_{exp}^{i}}{w_{theo}^{i}} \tag{7.2}$$

$maxdist$ is the maximal value of $dist$ used in the evaluation of the distributions. This is needed for numerical reasons. In our cases, the parameter $maxdist$ is chosen to be 20 bp, since the distance 20 occurs very rarely in DNA sequences. For example, if the probability p is 0.5, the quantity equals 10^{-6}.

The Kullback measure has some important properties: First, this quantity becomes equal to zero only if the compared distributions are indistinguishable; and second, the Kullback measure is a convex function, i.e. it never takes a negative value.

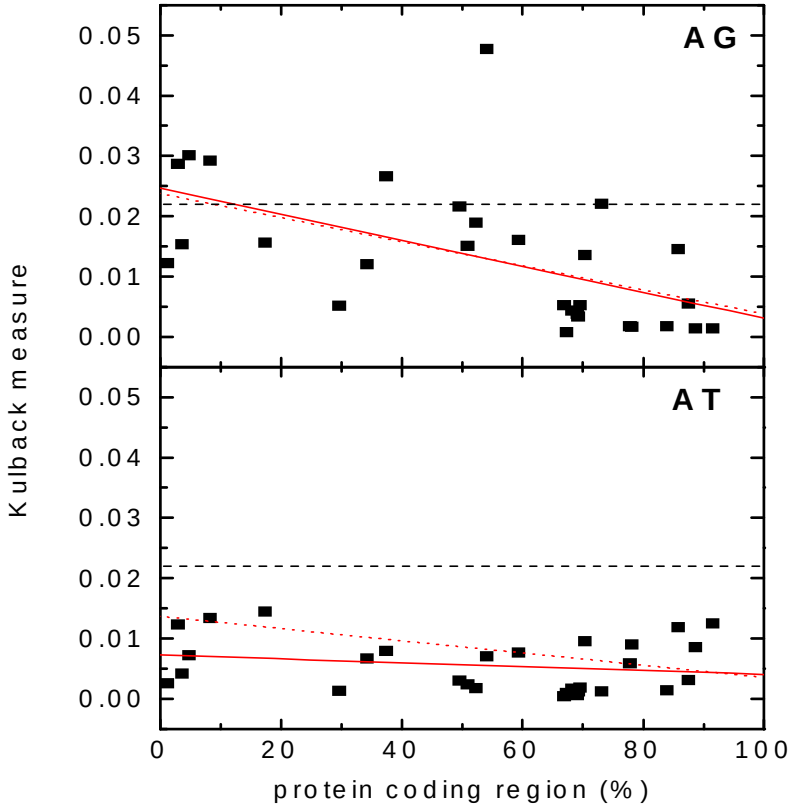

Figure 7.1: Dependence of the Kullback measure on the protein coding part of a sequence. Shown are the base combination AG and AT (data points and solid lines). The dotted lines correspond to the complementary base combinations CT and CG. The data points of these combination are omitted for clarity. Lines are fitted to the data points according to the standard linear regression method. The dashed line is the limit of an artificial sequence with constant base composition.

7.2 Results

In the present analysis of DNA sequences, the events of interest have to be defined at first, i.e. the specific base (or specific bases) between which the distribution of distances $dist$ is to be determined. These events will be referred to as *base combinations*. E.g., base combination (A, T) means that the distribution of the distances $dist$ between the bases adenine and/or thymine in a given DNA sequence are investigated.

The results, presented in fig. 7.1, deal with the four base combinations (A, T), (C, G), (C, T) and (A, G). With guidance of the previous result of section 3.2.4,

the numerical values of the quantity Ku are plotted against the coding parts of the investigated natural DNA sequences. At first sight, it is conspicuous that there are two groups of very similar base combinations. Namely, one group contains the base combinations (C, T) and (A, G), and the other group contains the base combinations (A, T) and (C, G). During further investigation, the data were fitted according to the standard linear regression method, revealing the absence of a significant slope at base combination (A, T) and (C, G).

Another interesting feature of fig. 7.1 is related with the Kullback measure for almost completely *non-coding* sequences. For the base combinations (C, T) and (A, G), this value is $Ku \approx 21 \times 10^{-3}$. Moreover, this value is almost equal to the Ku-values of the associated random control sequences (with $Ku = 22 \times 10^{-3}$) having a constant base composition along it (dashed lines in fig. 7.1). This Ku-value of random sequences appears to be due to 'finite size effects', i.e. due to the finite length of the random sequences (being typically around 50,000 bp in our present calculations). On the other hand, with respect to the base combinations (A, T) and (C, G), which represent base pairs with two and three H-bonds, the corresponding Ku-values of natural DNA sequences appear to be substantially different from those of the artificial random sequences.

Furthermore, it may be interesting to note that all base combinations mentioned above exhibit almost equal Ku-values in the cases of DNA sequences with a very high coding part.

In Sumary, the Kullback measure of almost non-coding sequences for the base combinations (C, T) and (A,G) (i.e. the distribution of pyrimidines and purines in the base pair sequence) is virtually the same as the Kullback measure of the 'random' control sequence. On the other hand, the Kullback measure of almost non-coding sequences for the base combinations (A, T) and (G, C) (i.e. the distribution of two and three H-Bonds in the base pair sequence) is significantly different form the Kullback measure of the 'random' control sequence. Taking these findings into account, the following *speculation* may be considered: If there is any additional information in non-coding sequences (unknown up to now) at all, then it is more probably coded with the aid of the distances $dist$ between DNA base pairs containing two (or three) hydrogen bonds, than with the aid of the distribution of pyrimidines and purines in the sequence.

8 A Statistical Test of Protein Coding DNA

The statistical analyses investigated thus far in the last chapters aimed at distinguishing coding and non-coding parts of DNA. Therefore, a new interpretation, called fractality or language, was imposed on the part of DNA, which is considered to be non-coding or silent. On the other hand, Dreismann and Seifert [76] investigated the protein coding part of DNA in different species, respecting the known codon structure. Specifically, the $4^2 = 16$ pairs of consecutive DNA bases in the 5'-3' direction of the translation process were investigated. Applying the hypotheses that two very close protons (or H-bonds) along the axis of the DNA helix may be quantum entangled or correlated for short times, the 16 pairs are divided in three groups Q, N and R:

- The first group, called Q, contains the nucleotide pairs with quantum correlated protons.
- Each member of the second group (which is called N) either does not contain any quantum correlated protons, or it contains less quantum correlated protons than the associated pair of the group Q. E.g., AG contains two quantum correlated protons whereas GA contains none.
- The remaining nucleotide pairs belong to the third group, called R.

In short:

Group Q: AG, TG, TA, GC, CA and CT.

Group N: GA, GT, AT, CG, AC and TC.

Group R: AA, TT, CC and GG.

It should be emphasized that all four DNA nucleotides A, T, C and G occur in Q as well as in N three times, i.e. they are distributed with equal statistical weight in both groups. The introduction of the third group R preserves this equal distribution.

Another point to mention here is that the pair CG contains one quantum correlated pair of protons, and it could therefore be attached to group Q. But the CG pair is found in group N since the associated pair GC contains two quantum correlated proton pairs. See ref. [76] for more details.

8.1 Organisms

Several sequences, assigned in six groups of different organisms containing 10 long coding sequences each, have been investigated. The first group contains sequences of the bacterium *Haemophilus influenzae* which is the first organism being sequenced completely [77]. Second, the well known bacteriophage λ has been chosen. As a representative of plants the chromosome III of *Saccharomyces cerevisiae* (yeast), the first chromosome completely sequenced [78], has been chosen. The sequence of *Caenorhabditis elegans* [79] represents the invertebrates. In the fifth group are included coding sequences of vertebrate animals; the chicken, the rat, the mouse and the gorilla. In the last group some sequences of the human genome were selected. (According to specific biological viewpoints, one may combine the last two groups to a single one, or split them into more groups.)

8.2 Results

The first investigation of the mentioned coding DNA sequences is the most simple one: count and classify all (L-1) pairs of consecutive bases in the considered gene with a length of L bases. Now, let Q_{all} be the number of pairs belonging to group Q, N_{all} be the number of pairs of group N and R_{all} be the number of pairs of group R, all over the considered ten sequences in one organism group. F_{all} is the ratio

$$F_{all} = \frac{Q_{all}}{N_{all}} \tag{8.1}$$

This ratio F_{all} is independent of the overall base composition because the four nucleotides are evenly distributed among the numerator and the denumerator of this fraction. Only R_{all} is influenced by the base composition. If the considered possible quantum correlations do not influence the primary structure of DNA-sequences, it is expected that F_{all} is equal to 1, within small deviations. On the contrary, if it appears that $F_{all} > 1$, then it may be speculated that the enhanced quantum correlated dimers of the group Q could represent some - thus far unknown - evolutionary advantage with respect to those of the group N. Simply formulated: plain statistics require $F_{all} \approx 1$.

The results of the numerical investigations, presented in table 8.1, clearly show strong deviations from the stochastic average $F_{all} \approx 1$: While the sequences of *H. influenzae*, the bacteriophage λ and *S.cerevisiae* behave as conventionally expected, i.e. $F_{all} \approx 1$, the sequences of *C.elegans* contain more pairs belonging to group N than Q; i.e. $F_{all} < 1$. On the other hand, the sequences of the vertebrates and of the human genome contain about 30% and 26%, respectively, more pairs of group Q than N. Thus, these results show that DNA of different organisms contain a

different amount of quantum correlated protons. The possible quantum correlations seem to be more important in DNA of 'higher' animals (vertebrates) and the human genome than in the DNA of other organisms.

To give this striking effect more evidence, the biological purpose of a coding DNA sequence was taken into account. The protein coding DNA sequences were considered as sequences of nucleotide triplets, called codons, coding one amino acid each. The three nucleotides of any codon are denoted by 5'-$B_1B_2B_3$-3'. The occurrences of the nucleotide pairs of the patterns B_1B_2, B_2B_3 and B_3B_1 (always in the standard 5'-3'-direction) belonging to the three groups Q, N and R were counted separately. To mark the site of a counted nucleotide pair, an index is used: E.g. Q_{23} denotes the number of nucleotide pairs which are located at bases B_2 and B_3 of a codon and belong to the group Q found in the considered genes. Therefore, it follows that Q_{all}=Q_{12}+Q_{23}+Q_{31}. Note that the pair B_3B_1 is located between two adjacent codons whereas the other two sites (B_1B_2 and B_2B_3) are located within only one codon.

The fractions F_{nm} with $nm = 12$, 23 or 31 are defined in analogy to Eq. 8.1:

$$F_{nm} = \frac{Q_{nm}}{N_{nm}} \tag{8.2}$$

According to the well-known 'degeneracy' of the genetic code (see textbooks, e.g. [80]) the third nucleotide B_3 of a codon is not uniquely determined. In contrast, the distributions of the nucleotides B_1 and B_2 of a codon are almost exclusively determined by the amino acid sequence. The results are summarized in table 8.1.

Note that (a) all averages of F_{12} are smaller than 1, (b) that the averages of F_{23} scatter about 1, except for the genes of yeast, the vertebrates and of the human genome whose averages are 1.18, 1.43 and 1.37 respectively and that (c) all averages over the F_{31}-values are larger than 1. All groups of DNA sequences exhibit, on average, larger values for F_{31} than for F_{12} or F_{23}. So it is obvious that on the sites containing B_3 more quantum correlated nucleotide pairs are found than on the site B_1B_2. Furthermore the trend of the first calculations could be mostly confirmed although more biological information was taken into account. The first calculations were made just by counting the base pair dimers without considering the nature of codons. Now the known triplet structure was taken into account and the effects do not vanish. In the next step the statistical weight of biological information is increased.

To improve the statistical significance of the above findings, a possible bias of the above results due to the base composition of the considered sequence has to be eliminated. Therefore, the natural DNA sequences were compared to artificial DNA sequences being associated with the natural ones, i.e. the artificial sequences

have the same base composition as the native sequences up to negligible finite size effects. For a detailed description of the numerical procedure and different concrete applications, see refs. [56, 57, 81].

A hundred different artificial sequences per native DNA sequence were created. The number of pairs belonging to the family Q were counted at the three mentioned locations of each artificial sequence and averaged over the different artificial sequences to give values called Q^*_{12}, Q^*_{23} and Q^*_{31}. (The asterisks indicate that these quantities were obtained from artificial sequences.) In other words, these averages quantify the appearance of base pairs with the 'property Q' in artificial sequences having the same base composition as the natural sequence. To quantify the 'differences' of the natural DNA sequences from the artificial ones the 'relative deviation' is defined:

$$\Delta F_{nm} = \frac{Q_{nm} - Q^*_{nm}}{Q^*_{nm}} \tag{8.3}$$

with nm=12, 23 or 31. The results are summarized in table 8.1.

First, all averages of ΔF_{12} are negative, with the human genes providing the less negative values. Second, some of the averages ΔF_{23} are positive (bacteriophage λ , yeast, vertebrate genes and human genes), others are negative (*H.influenzae* and *C. elegans*). In other words, at codon site B_2B_3 it seems to be dependent on the organism whether there are more quantum correlated pairs in the natural or in the artificial DNA-sequence. Third, all ΔF_{31}-values are positive. This means that in the natural DNA-sequences the quantum correlated pairs at codon site B_1B_2 are less than randomly expected, while the largest amount of quantum correlated pairs is located at B_3B_1. This result is consistent with the previous results, but a statistical more significant method was used. It still holds: The DNA of 'higher' organisms seem to contain more quantum correlated pairs than the DNA of other organisms.

8.3 Conclusion

The applied model is based on the 'working hypothesis' that the investigated quantum effect preferably occurs in pairs of consecutive base pairs, in which the H-bonds of the two base pairs may approach each other as much as possible. This assumption is motivated by the many-body interactions of condensed matter at ambient conditions, which are strongly attenuated by the decoherence effect (i.e., by a factor of $\exp\{-const(\Delta x)^2\}$) with increasing distance Δx of the two particles under consideration [23, 82]. It may be stressed here, that the considered effect does not concern idealized systems like ideal gases or perfect crystals.

Based on the above assumption, it is possible to differentiate between specific dimers of adjacent base pairs containing an enhanced number of possibly quantum entangled H-bonds, which define group Q, and the 'associated' dimers containing less or no such H-bonds, which define group N.

The quantitative analysis of the protein coding DNA sequences of various living organisms revealed the frequencies of occurrence of the groups Q and N. Most striking is the result that in protein coding DNA sequences the base pair dimer originating from the third base of a codon and the first base of the following codon is preferably chosen from group Q, which is assumed to have more quantum entangled H-bonds than group N.

In short, the considered quantum entanglement effect seem to appear preferred between two codons, and may form some linkage between these codons. Recall the well-known degeneracy of the genetic code, which is nowadays mainly related to the 'free choice' of the third base of a codon, corresponding to a specific amino acid. In this context it was necessary to take care of the correct reading frame. The above finding implies that the choice of the third base is not completely free, but it tends to increase the number of quantum entangled H-bonds between the third base of a codon and the first base of the following codon. This finding may indicate that quantum entanglement plays a specific, thus far unknown, role in the structure and/or dynamics of protein coding DNA sequences.

The magnitude of the effect is tremendous. Concentrating on the DNA of vertebrates (including human DNA) a ratio $F_{31}=Q_{31}/N_{31}$ of base pair dimers 'with' and 'without' quantum entangled H-bonds is about 1.8, i.e. about 80% larger than expected by simple statistics. Furthermore, the result has a high statistical significance. About 20000 codons were investigated, which indicate a statistical error of about 2% in F_{31}.

To give further support to the significance of this finding, the DNA sequences were compared to artificial sequences having the same length and same global base pair composition. The direct comparison between the natural and artificial sequences clearly shows a still increased amount of quantum entangled base pair dimers (about 25%) 'between adjacent codons' in the natural sequences. But, the relevance of this comparison should not be overestimated, since the artificial sequences resemble the natural DNA with respect to the base pair composition and length, but not with respect to the information content, i.e. the artificial sequences do not code for the same protein. Nevertheless, this comparison clearly demonstrates that the obtained result is not just an artifact of a biased base pair composition of the natural DNA sequences.

Analyzing the result from the viewpoint of traditional quantum chemistry, one may assume the main result (say, $F_{all} > 1$ and $F_{31} > 1$) is trivially caused by a difference in the stacking energies of the base pair dimers in group Q and N. Then the averaged stacking energies of the base pair dimers of group Q would

Table 8.1 Fractions F_{all} (see Eq. 8.1), F_{nm} with nm=12,23,31 (see Eq. 8.2) as well as their relative deviations ΔF_{nm} (see Eq. 8.3). See text and ref [76] for detailed information.

Organism	length [bp]	F_{all}	F_{12}	F_{23}	F_{31}	ΔF_{12} [%]	ΔF_{23} [%]	ΔF_{31} [%]
H.influenzae	37,959	1.03	0.63	1.06	1.60	-29.52	-4.57	20.87
Bacteriophage λ	16,923	1.03	0.79	1.05	1.38	-10.81	1.39	16.76
S.cerevisiae	37,755	1.01	0.66	1.18	1.30	-27.18	4.34	13.22
C.elegans	47,379	0.91	0.74	0.91	1.12	-22.50	-5.01	10.48
Vertebrates	37,716	1.30	0.84	1.43	1.89	-12.02	17.88	28.73
Homo sapiens	21,432	1.26	0.97	1.37	1.72	-6.11	14.38	24.09

be expected to be more negative, i.e. more attractive, than those of group N. However, testing this assumption on the base of data available in the literature [83–90] a clear contradiction arises, even though the applied methods are limited. In almost all investigations [83–88] the nucleotide pair of group N appear to be more favourable than those of group Q. Therefore, it would be expected that $F_{all} < 1$, but not $F_{all} > 1$. The data of refs. [89, 90] suggest no preference for either of the two groups. In summary, the differences of the stacking energies do not seem to be able to account for $F_{all} > 1$.

At last, it may speculated if the effect under consideration does really cause some 'evolutionary advantage' [76]. If so, nature would surely have made already 'use' of it. This may be manifested in the 'specific features' of the protein coding DNA investigated. The large value of F_{31} compared to the statistical estimates seems to be of particular biological interest. I.e. the third base of a codon B_3 may be used to 'store' a hitherto unknown biological information; e.g. it may be chosen properly in order to increase the quantum correlation stability of DNA along the helical axis or the entangled H-bonds provide a marker which could facilitate the transcription or the repair of DNA after some damage. The latter would also suggest a possible explanation of the well-known phenomenon of varying codon usage [91].

9 Remarks

One should not forget that DNA, the carrier of genetic information, is an extremely complex molecular system being the product of about three billion years of evolution of life. Therefore it should not be surprising that (relatively) simple statistical analyses (like the aforementioned 'long-range correlations', 'entropy and redundancy', 'Zipf distribution' etc.) have thus far not been able to reveal, and demonstrate reliably, any new 'biological information', neither in coding nor in non-coding DNA sequences. On the other hand, since at present a large number of DNA sequences of living organisms is known (and easily accessible through the DNA data bases), it is conceivable that true interdisciplinary efforts in the research work concerning the structure, organization and dynamics of DNA may be more successful in the near future.

Part II
Quantum Mechanical Analyses of DNA

In this part the double proton transfers in adenine-thymine (AT), guanine-cytosine (GC) and the $\kappa\chi$[3] base pairs are studied [92]. The results of a wide variety of quantum chemical methods are presented and discussed. Furthermore, based on the quantum chemical results, a short investigation of the dynamics of the proton in the base pairs is presented.

10 General remarks

In the frame of the Born-Oppenheimer approximation a complete description of the dynamics of the nuclei in a molecular system would require the full knowledge of the hypersurface, which involves the solution of the electronic Schrödinger equation for all possible configurations of the nuclei. Unfortunately, the analytic form of the hypersurfaces of molecules larger than two nuclei and one electron are not known up till now. Therefore the hypersurface has to be scanned in a point by point manner [93]. E.g. an AT base pair is formed by 30 atoms with 84 degrees of freedom. The calculation of only 10 points on each intersect of the hypersurface would imply the iterative solution of 10^{84} electronic Schrödinger equations. Even if each aforementioned solution would only take one second, all of them would take about 10^{76} years. In comparison, the universe is only about 10^{9} to 10^{10} years old.

Despite of this estimation an investigation of the dynamics of the nuclei is not a priori impossible, but one has to restrict oneself to the so-called 'relevant' parts of the hypersurface. This means, all possible pathways across the hypersurface describing the reaction are not equally important. The most important paths are usually those which couple the movement of the nuclei, to cross the saddle point with the lowest potential energy between the educt and product minimum on the hypersurface. For the estimation of molecular movements it is often sufficient to determine the transition state (TS) and the two minima. Therefore, to optimize the accuracy of the result and the computational effort, two different approaches are used to investigate the proton dynamics of the H-bonds in the base pair AT, GC and $\kappa\chi$:

1. The possible concerted double proton transfers in the AT, GC and $\kappa\chi$ base

[3] $\kappa\chi$ is an artificially designed and synthesized base pair described in detail in section 10.2

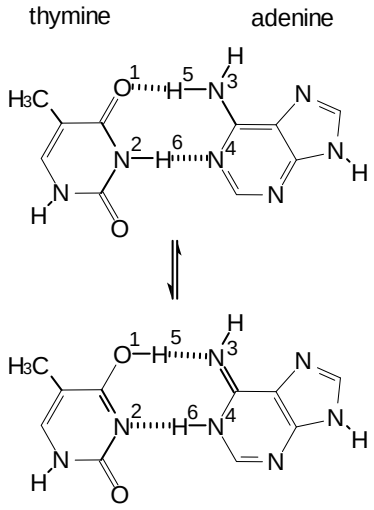

Figure 10.1: Schematic structures of the AT base pair (upper half) and its associated tautomer (bottom half) after the DPT reaction. The protons H_5 and H_6 are shifted along the lines N_3-O_1 and N_2-N_6 respectively during the potential energy scan. H-bonds are denoted by striped lines.

pairs are studied by the means of the so-called 'frozen core' approximation, i.e. all atom positions are fixed and only the relevant protons are shifted along the connecting line of the associated heavy atoms (see section 11) The method is analogous to the one used by Kong et al. [94].

2. The stationary points of the hypersurface are determined by a complete geometry optimization of the minima on the hypersurface and the transition state. Furthermore the vibrational spectra in the harmonic approximation are discussed (cf. section 12).

10.1 Proton tunnelling in DNA base pairs and its biological significance

Since the discovery of the structure of DNA, it has been conjectured that tautomeric forms of the bases A, T, C and G might cause 'errors' during DNA replication and associated processes. Already in 1953 Watson and Crick stated: 'our model suggests possible explanations for a number of other phenomena. For example, spontaneous mutation may be due to a base occasionally occurring in one of its less likely tautomeric forms' (see [1]).

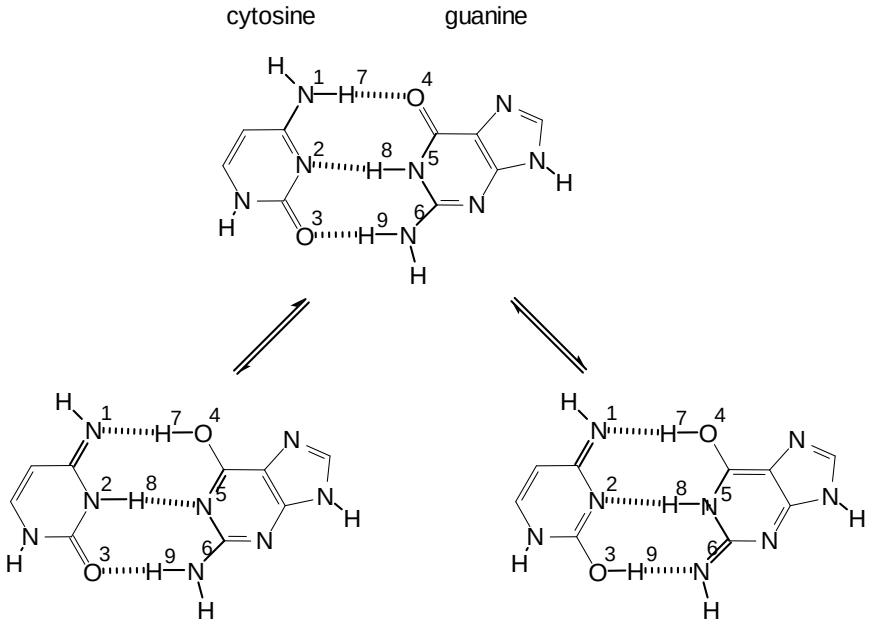

Figure 10.2: Depicted are the two possible DPT of the GC base pair. The left branch shows the first DPT involving H_7 and H_8, which are shifted along the lines N_1-O_4 and N_5-N_2, respectively, leading to tautomer 1. The right branch shows the second possible DPT (tautomer 2) involving the movement of H_8 and H_9 along the lines N_1-O_4 and N_6-O_3. H-bonds are denoted by striped lines.

Tautomeric forms of DNA bases are crucial for Löwdin's 'double proton transfer' model of mutagenesis, which has been studied for many years [40, 95]. A (coupled) double proton transfer in an AT, GC or $\kappa\chi$ base pair corresponds to the following (cf. figs. 10.1, 10.2 and 10.3): If one proton of the two H-bonds in the AT (or of the three H-bonds in the GC base pair) moves from its equilibrium position near its N-atom along the line of the H-bond, to the lone electron pair of its opposite O-atom, then this is likely to induce the reverse motion of a second proton in another H-bond of the same base pair. The latter movement maintains the gross electric neutrality of the base pair [40]. In the GC base pair two H-atoms of the three H-bonds 'belong' to guanine, which, together with the H-atom 'belonging' to cytosine (see fig. 10.2), may undergo the coupled proton transfer under consideration. Therefore, two different double well potentials exist, one for each double proton transfer.

Early quantum chemical calculations showed that the specific double proton transfer in DNA base pairs indicated above may lead to double-well potential energy profiles (in short 'potentials' or 'PESs') for the two protons involved. The

derived potentials support the formation of some short lived tautomeric forms of the bases. As Löwdin pointed out, the protons are not classical particles but 'wave packets' obeying the laws of quantum theory and thus they may be subject to the well-known tunnel effect in the double-well potentials mentioned above. This effect implies that the genetic code cannot be 100% stable, which furthermore means that this proton transfer of about 1 Å may be one of the driving forces in the evolution of living organisms on earth [40].

In this context, it should be mentioned that Clementi et al. carried out an *ab initio* SCF calculation for the possible tunnelling of *one* proton in the GC base pair. They found that the *single* proton transfer (which forms an ion pair) shows a single well energy profile characterized by a monotonically increasing energy function [96].

More recent calculations (at the *ab initio* SCF level) on the double proton transfer, however, support the existence of the aforementioned double-well potentials [94]. Interesting enough to note that these calculations exhibit far more asymmetric double-well potentials than the previously reported work [40]. This means that the structure of the base pairs in DNA appears to be more stable than previously believed[4]. As a consequence, the experimentally determined spontaneous mutation rates in DNA, as cited in [40], cannot be explained by these calculation in a quantitative way.

Nevertheless, the above results support qualitatively Löwdin's hypothesis that certain tautomeric forms of the bases may result from double proton tunnelling in the Watson-Crick type base pairs in the DNA double helix. Additionally, the results of Kong et al. [94] indicate that the possibility of an error (or mutation) in the genetic code replication occurs more easily in a GC base pair than in an AT base pair, since the barrier in the double-well potential of the GC base pair is considerably lower than that in the AT base pair. In other words, the equilibrium concentrations of the tautomeric forms is expected to be greater in the case of the GC base pair than in the case of the AT base pair. It has been speculated that this observation might offer a possible explanation for the larger relative AT content in higher organisms [40].

10.2 The artificial $\kappa\chi$ base pair – an extension of the genetic alphabet

The geometry of a Watson-Crick type base pair can accommodate several mutually exclusive hydrogen bonding schemes (see figs. 10.1, 10.2 and 10.3). Each of them is defined by the distribution of H-bonds (proton-donor and -acceptor) on the

[4]The most recent calculations on this topic which do not deal with PES scans, but optimize the structures, are surveyed in section 12

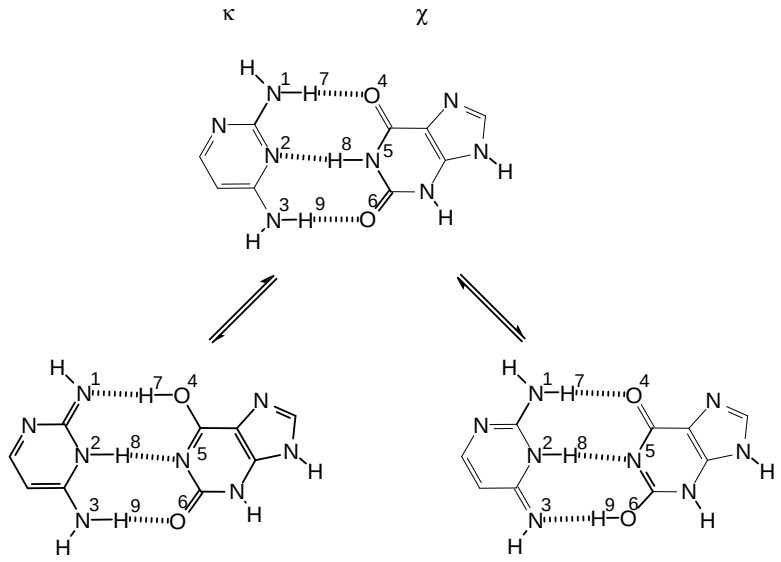

Figure 10.3: Shown are the two possible DPTs of the $\kappa\chi$ base pair. The left branch shows the first double proton transfer involving the movement H_7 and H_8 along the lines N_1-O_4 and N_5-N_2 to form tautomer 1. The right branch shows the second possible DPT (tautomer 2) during which H_7 and H_9 are shifted along N_5-N_2 and N_3-O_6.H-bonds are denoted by striped lines.

purine and pyrimidine rings. Nature, however, uses only two of these schemes, known as AT and GC base pair.

Recently a new Watson-Crick type base pair, called $\kappa\chi$ (see fig. 10.3), exhibiting a different H-bond pattern from that in the AT and GC base pair, was designed and synthesized [97–99]. Furthermore, this base pair was successfully incorporated into duplex DNA and RNA by adequate polymerases. Additional melting experiments with several oligonucleotides showed that duplexes containing a $\kappa\chi$ base pair are only slightly less stable than duplexes containing only AT and GC base pairs. Moreover, duplexes containing the new base pair appear to be considerably more stable than those containing mismatches involving the new bases, which in turn have melting temperatures similar to duplexes containing mismatches of natural bases [97].

Thus this pioneering biochemical work demonstrated the feasibility of expanding the genetic alphabet by increasing the number of letters that can be incorporated into nucleotides enzymatically by template directed polymerisation. Among other consequences, it has been pointed out [97] that new RNA molecules based also on the new bases have the potential for an even increased catalytic power.

11 Investigations in the 'frozen core' approximation

11.1 Methods

The potential energy surfaces of the double proton transfer (DPT) reaction in the AT, GC and the $\kappa\chi$ base pair were examined by means of *ab initio* methods. The geometry of the AT and GC base pair was taken from ref. [100]. The positions of the protons, which cannot be resolved by X-ray cristallography, are determined by the minimum of the B3LYP/6-31G** level of theory. B3LYP denotes the method of Density Functional Theory (DFT) using Becke's three parameter exchange functional with the Lee, Yang and Parr correlation function. The geometry of $\kappa\chi$, which is not known by experiment, was determined by the minimum of the B3LYP method with split valence basis set 6-31G**. This method gave the best results when applied on earlier geometry calculations [92].

Among the applied methods were the simple Hartree-Fock approximation with basis sets STO-3G and 6-31G**, the second order Møller-Plesset perturbation theory [101], and the B3LYP [102] method of approximate density functional theory [103]. For both methods estimating the electron correlation energy the split valence basis set 6-31G** was applied. The 6-31G** basis set includes functions of s- and p-type on H-atoms and s-, p-, and d-type on C-, N-, and O-atoms. Additionally the B3LYP/D95 level of theory was investigated in case of the $\kappa\chi$ base pair to get an impression of the influence of the polarization functions.

The potential energy surface of the double proton transfer was scanned step by step in distances of 0.1 Å in the so-called frozen core approximation along each reaction coordinate, i.e. single point energy calculations are performed on fixed molecule geometries. Two H-bonds are participating in each reaction coordinate, which are synchronously elongated along the connecting line of the two heavy atoms of each H-bond. The reaction coordinate in the AT base pair is defined by the bonds N_3-H_5 and N_2-H_6, which are stretched along the lines N_3-O_1 and N_2-N_4 (see fig. 10.1). In the GC base pair two possible tautomer reactions exist. The first reaction coordinate, leading to tautomer 1 of GC, is formed by the bonds N_1-H_7 and N_5-H_8 which are elongated along the lines N_1-O_4 and N_5-N_2. The second reaction coordinate, with the product tautomer 2 of GC, involves the elongation of the bonds N_1-H_7 and N_6-H_9 along N_1-O_4 and N_6-O_3 (cf. fig. 10.2). Finally, in case of the $\kappa\chi$ base pair also two reaction coordinates exist, which are defined by N_1-H_7 and N_5-H_8 of $\kappa\chi$ elongated along N_1-O_4 and N_5-N_2, similar to the first reaction coordinate of GC, as well as N_5-H_8 and N_3-H_9 elongated along N_5-N_2 and N_3-O_6 (see fig. 10.3). All calculations have been carried out with GAUSSIAN 94 program suite [104].

Table 11.1 Relative potential energies of the DPT in the AT base pair in kcal mol^{-1} (see also fig. 11.1). In the head of the table the applied method and basis set are given along with the energy of the absolute minimum. The DPT reaction coordinate (DPTRC) is measured in Å. Its definition can be found in the legend of fig. 10.1 and the text.

method:	HF	HF	B3LYP	MP2
basis set:	STO-3G	6-31G**	6-31G**	6-31G**
abs.min. (H)	-904.28820	-916.05874	-921.47742	-918.86478
DPTRC(Å)				
0.8	72.6	47.2	54.9	54.6
0.9	14.6	6.2	10.5	9.9
1.0	0.0	0.0	0.0	0.0
1.1	8.5	11.1	6.0	7.2
1.2	28.0	28.8	18.3	20.9
1.3	49.3	46.1	30.7	34.5
1.4	65.6	58.1	39.5	44.2
1.5	72.7	62.2	43.0	48.1
1.6	70.0	58.6	41.5	46.3
1.7	62.1	50.2	37.2	41.3
1.8	57.8	43.3	35.5	38.4
1.9	70.2	47.9	45.2	47.4
2.0	118.4	80.8	82.5	84.7
2.1	234.4	170.6	175.5	178.9

11.2 Results

The calculated points of the PES of the all DPTs are given for all different *ab initio* methods in tabs. 11.1–11.5. The corresponding graphs are plotted in the figs. 11.1–11.5. In order to compare all PESs with each other the energies were normalized to the energy of the global minimum of each method applied, which is by definition 0.0 kcal mol^{-1}. All structures have been calculated in the 'frozen core' approximation, i.e. all atoms had fixed positions.

11.2.1 AT base pair

As a first relation to the literature [94] the (now minimalistic) HF/STO-3G level of theory has been applied. The results of the calculations are in good agreement with ref. [94] and show an asymmetric double-well in the PES. Possible minor deviations (below 2 kcal mol^{-1}) in the potential energy are due to the positioning of the hydrogen atoms. It should remembered that the density functional theory, and

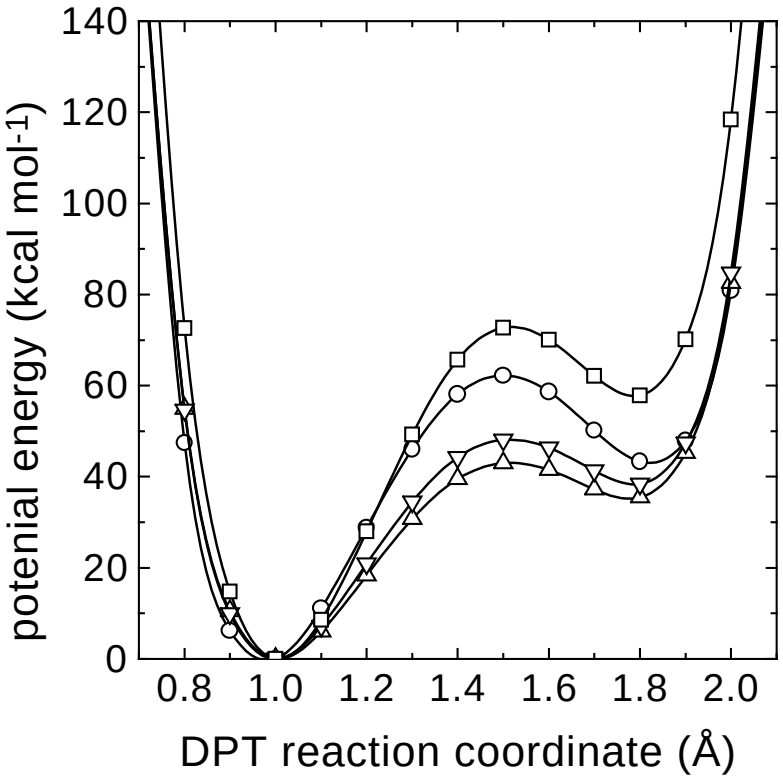

Figure 11.1: Depicted are the PESs of the DPT in the AT base pair at different levels of theory: □: HF/STO-3G; ○: HF/6-31G**; △: B3LYP/6-31G**; ▽: MP2/6-31G**. The reaction coordinate is specified in the legend of fig. 10.1 and the text. The solid lines are guides to the eye.

especially the B3LYP method, was not commonly applied in quantum chemistry at the time of ref. [94]. The results are summarized in fig. 11.1 and tab. 11.1.

Furthermore, all calculated PES of the DPT in the AT base pair have the asymmetric double-well feature. The minima of the PES reside at 1.0 Å and 1.8 Å, whereas the transition state (position of highest potential energy) is located at 1.5 Å independent of the theory applied. Basis set extension and inclusion of electron correlation energy reduce not only the potential energy of the second minimum from 57 kcal mol^{-1} to 36 kcal mol^{-1}, but affect mainly the transition state. The potential energy of the transition state is reduced by 10 kcal mol^{-1} on the basis set extension from the STO-3G minimal basis set to the 6-31G** split valence basis set. The inclusion of the electron correlation energy further reduces the potential energy of the transition state by 14-19 kcal mol^{-1} depending on the level of theory

Table 11.2 Relative potential energies of the DPT of the GC base pair involving H_7 and H_8 to form tautomer 1 in kcal mol^{-1} (see also fig. 11.2). In the head of the table the applied method and basis set are given along with the energy of the absolute minimum. The DPT reaction coordinate (DPTRC) is measured in Å. Its definition can be found in the legend of fig. 10.2 and the text.

method:	HF	HF	B3LYP	MP2
basis set:	STO-3G	6-31G**	6-31G**	6-31G**
abs.min. (H)	-920.01943	-932.06979	-937.54552	-934.88843
DPTRC(Å)				
0.8	71.8	46.2	53.4	53.4
0.9	14.6	5.7	9.8	9.4
1.0	0.0	0.0	0.0	0.0
1.1	8.5	11.7	6.6	7.7
1.2	27.7	30.4	19.8	22.1
1.3	48.8	49.1	33.2	36.7
1.4	64.8	62.8	43.2	47.4
1.5	70.8	68.2	47.6	52.1
1.6	65.6	64.7	46.3	50.2
1.7	53.0	54.2	40.4	43.3
1.8	40.6	41.8	33.8	35.5
1.9	38.9	35.7	33.5	34.1
2.0	63.4	48.9	51.8	51.9
2.1	138.1	103.2	110.1	110.8

applied (cf. fig. 11.1). The transition state has its minimal potential energy at the B3LYP/6-31G** level of theory, which is 43 kcal mol^{-1}.

11.2.2 GC base pair (tautomer 1)

The basic characteristic feature of the PES appears to be, as for all different PES under consideration, the asymmetric double-well potential. The two minima are located at about 1.0 Å and 1.9 Å, and the transition state can be found at about 1.5 Å, independent of the level of theory applied in the calculation. A summary of the data is given in fig. 11.2 and tab. 11.2.

In the frame of the Hartree-Fock calculations the substitution of the basis set STO-3G, used by Kong et al. [94], with the significantly larger split valence basis set 6-31G** does not change the relative energies remarkably. Regarding the correlation effects of the electrons within the MP2 and the DFT methods the impact on the energy difference of the transition state in each applied method was

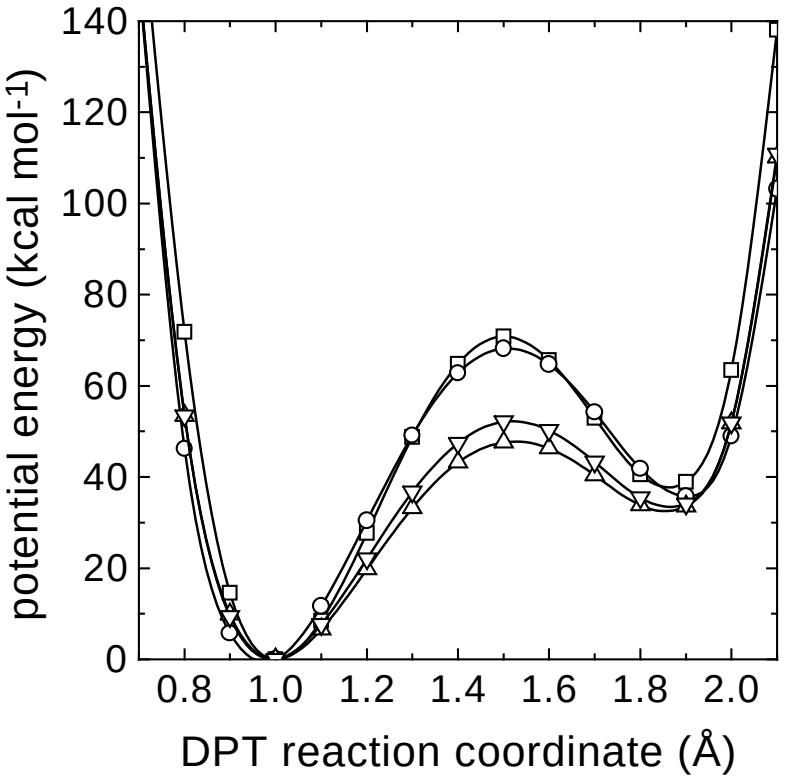

Figure 11.2: PES of the DPT involving H_7 and H_8 of the GC base pair to form tautomer 1 (see left branch of fig. 10.2) at different levels of theory: □: HF/STO-3G; ○: HF/6-31G**; △: B3LYP/6-31G**; ▽: MP2/6-31G**. The solid lines are guides to the eye. The definition of the reaction coordinate is given in the legend of fig. 10.2.

enormous, namely almost 20 kcal mol^{-1} compared to both minima. The energy difference between the global and the second minimum did not change much, whereas the energy difference between the second minimum and the transition state decreased, which implicates a smaller depth of the wells. The depth of the second well decreased from 32 kcal mol^{-1} to 14 kcal mol^{-1} according to the transition state.

11.2.3 GC base pair (tautomer 2)

The PES calculated for the second tautomer is completely different from the first one. The summarized results can be found in fig. 11.3 and tab. 11.3. Although the two minima calculated in the Hartree-Fock theory reside at 1.0 Å and 1.85 Å, the

Table 11.3 Relative potential energies of the DPT of the GC base pair involving H_7 and H_9 to form tautomer 2 in kcal mol^{-1} (see also fig. 11.3). In the head of the table the applied method and basis set are given along with the energy of the absolute minimum. The DPT reaction coordinate (DPTRC) is measured in Å and is defined in the legend of fig. 10.2 and the text.

method:	HF	HF	B3LYP	MP2
basis set:	STO-3G	6-31G**	6-31G**	6-31G**
abs.min. (H)	-920.01943	-932.06979	-937.54552	-934.88843
DPTRC(Å)				
0.8	75.9	49.8	56.8	56.1
0.9	15.8	6.8	10.9	10.0
1.0	0.0	0.0	0.0	0.0
1.1	8.2	11.7	6.9	8.3
1.2	28.1	31.3	21.1	24.2
1.3	51.0	51.8	36.3	41.2
1.4	70.3	68.6	48.9	55.2
1.5	81.8	78.3	57.1	63.9
1.6	83.6	79.8	59.8	66.8
1.7	78.2	74.6	58.6	64.8
1.8	72.4	67.1	56.8	61.9
1.9	78.0	66.2	62.0	66.1
2.0	112.6	86.9	88.3	92.0
2.1	205.2	155.5	161.6	165.9

distance between the first minimum (global minimum) and the transition state in the second tautomer is bigger (0.6 Å) than that in the first tautomer (0.5 Å). The larger basis set in the HF approximation widens the distance between the minima by approximately 0.1 Å, but this has little impact on the potential energy of the two minima and the transition state. The energy difference between the structures of the second minimum at 1.8 Å and the transition state at 1.6 Å equals only about 3-4 kcal mol^{-1} for the correlated methods.

11.2.4 $\kappa\chi$ base pair

Here a major result of the present investigation will be presented. Comparing fig. 11.4 and fig. 11.5 with each other it is evident that the PES of both tautomers are nearly identical at each level of theory applied, i.e. the difference vanishes (cf. fig. 11.6) independent of the theory applied using large basis sets. Only while applying the minimal basis in the Hartree-Fock theory the difference is slightly

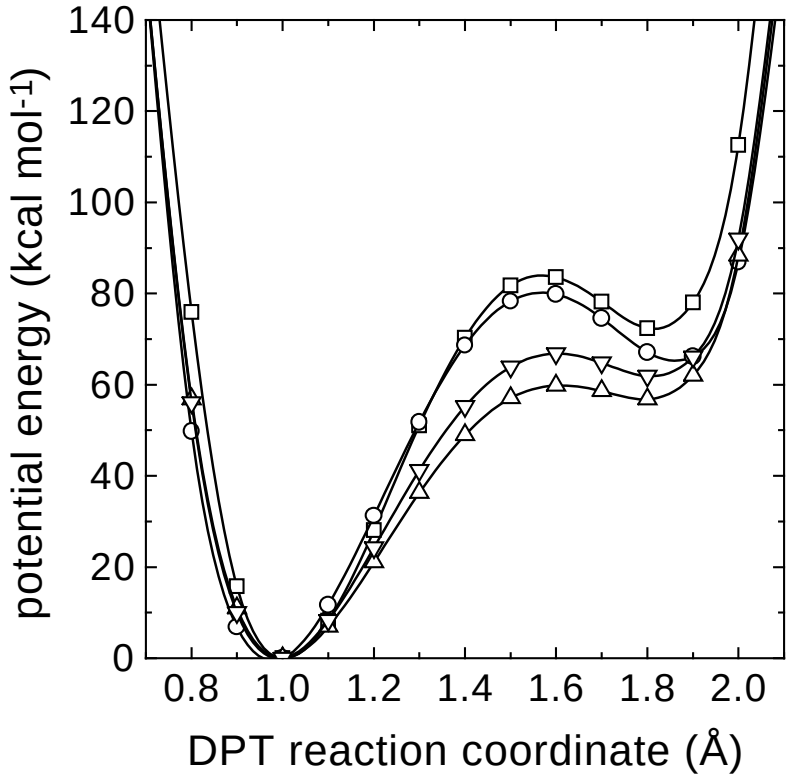

Figure 11.3: PES of the DPT shifting H_7 and H_9 to form the product tautomer 2 of the GC base pair (see right branch of fig. 10.2). The levels of theory investigated are as follows: □: HF/STO-3G; ◯: HF/6-31G**; △: B3LYP/6-31G**; ▽: MP2/6-31G**. The reaction coordinate is defined in the legend of fig. 10.2 and the solid lines are guides to the eye.

Table 11.4 Relative potential energies of the DPT of the $\kappa\chi$ base pair forming tautomer 1 in kcal mol^{-1} (cf. figs. 10.3 and 10.3). In the head of the table the applied method and basis set are given along with the energy of the absolute minimum. The DPT reaction coordinate (DPTRC) is measured in Å and is defined in the legend of fig. 10.3 and the text.

method:	HF	HF	B3LYP	B3LYP	MP2
basis set:	STO-3G	6-31G**	6-31G**	D95	6-31G**
abs.min. (H)	-920.046482	-931.612740	-937.546091	-937.696730	-932.066977
DPTRC(Å)					
0.8	74.0	49.3	56.4	58.1	56.3
0.9	15.0	6.7	10.8	11.5	10.3
1.0	0.0	0.0	0.0	0.0	0.0
1.1	8.9	11.4	6.3	6.0	7.5
1.2	29.5	30.2	19.7	19.1	22.3
1.3	53.1	49.4	33.7	33.0	37.8
1.4	73.1	64.3	44.8	44.0	50.0
1.5	84.6	71.8	50.7	49.9	56.7
1.6	85.5	70.8	51.1	50.2	57.1
1.7	78.0	63.0	47.0	46.0	52.5
1.8	68.6	52.5	41.8	40.8	46.2
1.9	67.4	47.0	41.7	41.0	45.1
2.0	89.2	58.4	57.9	57.8	60.9
2.1	156.8	106.8	110.2	111.2	113.6

deviating from zero. As will be seen later this may have biological consequences in the light of the arguments of ref. [105].

The minima reside at 1.0 Å and 1.9 Å, as well as the transition state at 1.55 Å. At the Hartree-Fock level of theory introducing a larger split valence basis set, i.e. 6-31G**, the energetic difference decreases between the transition state and the global minimum by about 13 kcal mol^{-1}; whereas between the second well and the global minimum the energetic difference decreases by about 20 kcal mol^{-1}. Thus the depth of the second well grows by about 8 kcal mol^{-1} on this basis set extension. Including electron correlation energy (e.g. with the MP2 method) has little or no effect on the locations of the minima and the transition state. There is also no sharp difference of the potential energies of the second well (about 2 kcal mol^{-1}) compared with the Hartree-Fock level of theory using the 6-31G** basis set. But the potential energy of the transition state decreased by nearly 14 kcal mol^{-1}, just as much as for the basis set extension on the Hartree-Fock level. Thus the second well flattens enormously, similar to the calculation of the GC base pair.

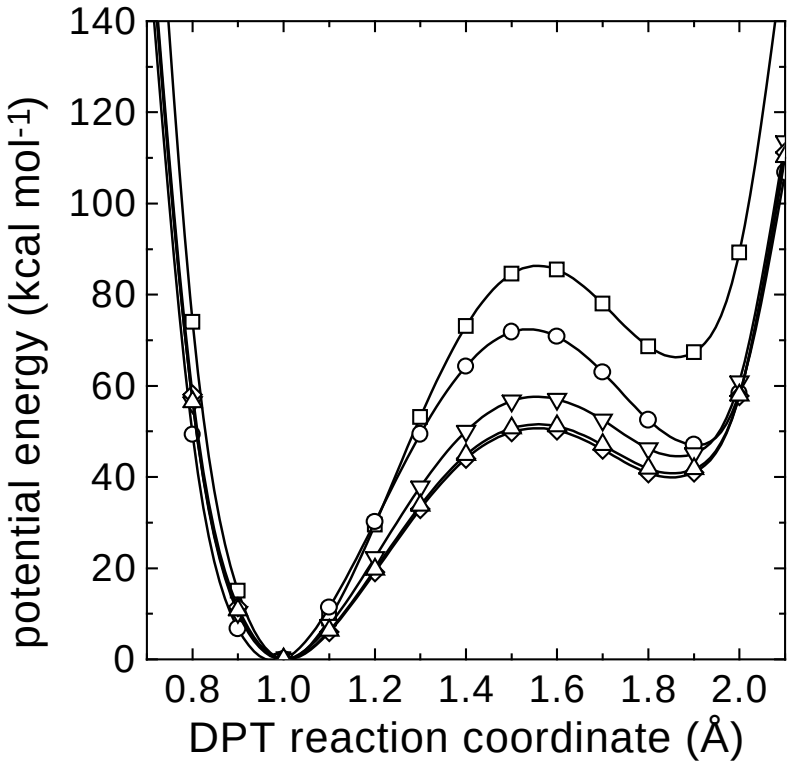

Figure 11.4: Depicted are the PESs of the DPT forming tautomer 1 of $\kappa\chi$ (see left branch of fig. 10.3) by shifting H_7 and H_8 along the reaction coordinate at different levels of theory: □: HF/STO-3G; ○: HF/6-31G**; △: B3LYP/6-31G**; ◇: B3LYP/D95; ▽: MP2/6-31G**. For the definition of the reaction coordinate see the legend of fig. 10.3. The solid lines are guides to the eye.

Regarding the empiric electron correlation energy, obtained by the B3LYP method, as compared with electron correlation energy of the MP2 method, an even larger decrease of the relative energy of the transition state to the global minimum (about 6 kcal mol^{-1}) is obtained, .

In order to test the influence of the polarization function of the 6-31G** basis set, the well known D95 basis set by Dunning [106] has been applied. This is a basis set with about the same number of functions but without the p-type functions on H-Atoms and the d-type functions on the heavier atoms. There is only little difference ($<$1 kcal mol^{-1}) between both applied basis sets, 6-31G** and D95, suggesting the number of basis functions to be more important for the current problem than their type.

Table 11.5 Relative potential energies of the DPT of the $\kappa\chi$ base pair forming tautomer 2 in kcal mol^{-1} (see also fig. 10.3 and 11.5). In the head of the table the applied method and basis set are given along with the energy of the absolute minimum. The DPT reaction coordinate (DPTRC) is measured in Å and is defined in the legend of fig. 10.3 and the text.

method:	HF	HF	B3LYP	B3LYP	MP2
basis set:	STO-3G	6-31G**	6-31G**	D95	6-31G**
abs.min. (H)	-920.046482	-931.612740	-937.546091	-937.696730	-932.066977
DPTRC(Å)					
0.7	211.9	159.3	168.2	171.2	169.6
0.8	74.3	49.4	56.5	58.2	56.3
0.9	15.1	6.7	10.8	11.5	10.2
1.0	0.0	0.0	0.0	0.0	0.0
1.1	8.9	11.4	6.4	6.0	7.6
1.2	29.6	30.2	19.7	19.1	22.5
1.3	53.3	49.5	33.7	33.0	38.1
1.4	73.4	64.3	44.6	43.9	50.4
1.5	85.0	71.7	50.4	49.6	57.2
1.6	85.9	70.5	50.6	49.8	57.7
1.7	78.4	62.5	46.4	45.5	53.2
1.8	69.1	51.9	41.0	40.2	47.1
1.9	68.3	46.4	40.9	40.4	46.4
2.0	90.6	58.0	57.4	57.5	62.3
2.1	158.9	106.7	110.0	111.2	115.5

11.3 Discussion

The calculations presented [92], employing large basis sets with and without polarization functions as well as an estimation of the electron correlation energy, are a major enhancement of the data provided by ref. [94]. In these calculations the concerted DPT in DNA base pairs is considered, which includes no energetically disfavoured ionic structures. The PESs of the AT and GC base pair have, in good agreement with ref. [94], the double well structure at all levels of theory applied. In fact the double well structure for tautomer 2 of the GC base pair (see right branch of fig. 10.2) becomes very weak on basis set extension and inclusion of the electron correlation energy, i.e. the second minimum is very flat, only 3 kcal mol^{-1} deep at the B3LYP/6-31G** level of theory.

Concentrating in a first attempt to explain this behaviour on the base pairs available to nature, i.e. AT and GC, it is found that the electron system, rearranged

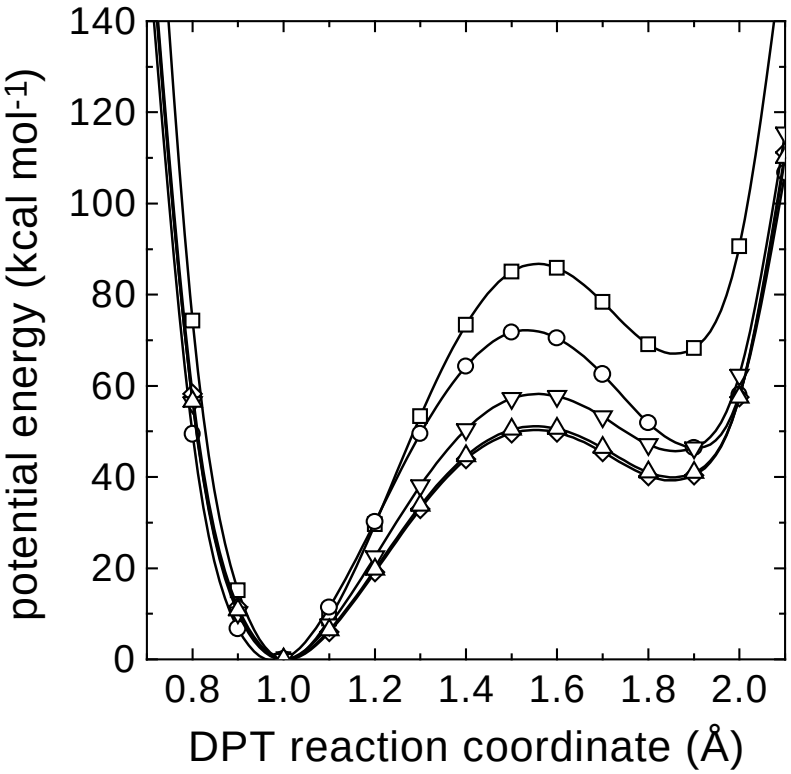

Figure 11.5: PES of the DPT involving H_8 and H_9 in the $\kappa\chi$ base pair leading to tautomer 2 as shown in the right branch of fig. 10.3). The level of theory applied are as follows: □: HF/STO-3G; ◯: HF/6-31G**; △: B3LYP/6-31G**; ◇: B3LYP/D95; ▽: MP2/6-31G**. For the definition of the reaction coordinate see the legend of fig.10.3. The solid lines are guides to the eye.

during the DPT reactions, is by far the largest for the DPT forming tautomer 2 of GC, i.e. 7 electron pairs (cf. fig. 10.2). In the other two cases (the DPT in the AT base pair and the DPT of GC forming tautomer 1) the rearranged electron system consists of only 4 electron pairs. If this is the explanation for the behaviour of tautomer 2 of the GC base pair, then it would be expected that the DPTs in the $\kappa\chi$ base pair rearrange 4 electron pairs as estimated by the energy of the second minimum of the PES. The DPT forming tautomer 1 of $\kappa\chi$ rearranges indeed 4 electron pairs, but, on the contrary, the DPT forming tautomer 2 rearranges 6 electron pairs (cf. fig. 10.3). Therefore, this reasoning ia a non-valid explanation.

Further investigations on the type of the involved H-bonds show that an explanation for the behaviour of the GC base pair might be given as follows: A

symmetric H-bond of type N-H-N and an asymmetric H-bond of type N-H-O are participating (i) in the DPT of the AT base pair, (ii) in the DPT of GC leading to tautomer 1 as the product and (iii) in both DPTs of the $\kappa\chi$ base pair. All these DPTs have a fairly stable second well with a depth between 7 kcal mol^{-1} and 9 kcal mol^{-1}. Only the DPT of GC forming tautomer 2 involves two asymmetric H-bonds of type N-H-O, reducing the depth of the second well to only 3 kcal mol^{-1}. This is also consistent with the fact, that the amino-ketone form of a molecule (found in a base pair) is more stable than its corresponding imino-enol form (found in a tautomer of a base pair). According to the numerical results, it seems to be possible to stabilize one imino-enol tautomer during the DPT reactions but not two. According to the aforementioned numerical results, the real existence of tautomer 2 of the GC base pair is doubtful. Therefore, further effort will be concentrated on tautomer 1 of GC.

Based on the double well structure of the PES for tautomer 1 of GC it is possible to adopt Löwdin's proton tunneling model for mutagenesis [40, 95, 107]. The same does also hold for the DPT of the AT base pair and both DPTs of the artificial $\kappa\chi$ base pair. More precisely, it is possible to consider in all cases (GC, AT and $\kappa\chi$ base pair) an "error mechanism" during the replication of the genetic code based on the double proton transfer forming the corresponding tautomer.

Here let us try a direct comparison of the tautomerisation reactions based on simple thermodynamics. First of all, the relative amount of each tautomer at any time, i.e. the equilibrium constant, is about the same in case of the GC and AT base pair, but considerably smaller for both DPTs of $\kappa\chi$. This is expressed in the energetic difference of the two minima of each individual PES, which is connected to the equilibrium constant. Explicitly this difference amounts to 34 kcal mol^{-1} and 36 kcal mol^{-1} in case of the GC and AT base pair, respectively, and 42 kcal mol^{-1} in case of the $\kappa\chi$ base pair at the B3LYP/6-31G** level of theory.

Concerning the dynamic properties of the different reactions, i.e. the rate constants, which are connected to the relative energy of the transition state, the picture is a little different. The relative energies of the transition states, which are connected to the rate constants k_x (with x=GC, AT or $\kappa\chi$) of the tautomer formation, are 48 kcal mol^{-1} for GC and 51 kcal mol^{-1} for $\kappa\chi$, which suggests the rate constants k_{GC} and $k_{\kappa\chi}$ to be of the same order. The rate constant k_{AT} will be considerably higher, which is expressed in a lower relative energy of the transition state, e.g. 43 kcal mol^{-1}. Investigating the rate constants k_x^{back} (with x=GC, AT or $\kappa\chi$) of the revers- or backreactions, $k_{\kappa\chi}^{back}$ and k_{AT}^{back} are similar, based on the energetic difference of the transition state and the second minimum, which is 9 kcal mol^{-1} and 7 kcal mol^{-1} for $\kappa\chi$ and AT respectively. The reformation of the GC base pair with a 14 kcal mol^{-1} barrier is slower.

In the light of Löwdin's hypothesis about mutagenesis this leads to a higher formation rate of the tautomer in the AT base pair as compared with the GC base

pair (i.e. $k_{AT} > k_{GC}$), but the tautomer of AT in turn lives not as long as the tautomer of GC does (i.e. $k_{AT}^{back} > k_{GC}^{back}$). On the average, there are approximately the same relative amounts of AT tautomers and GC tautomers at any time, which is expressed by similar equilibrium constants or the difference between the two minima of the PES. This suggests that the occurrence of an error or point mutation is equally probable in the AT and GC base pair. Kong et al. [94] had found a higher relative energy of the transition state of the AT base pair, a result which was due the low level of theory applied, and was enforced by the restriction of computer resources at that time.

The kinetics of the $\kappa\chi$ base pair is somehow similar to the kinetics of either the AT and GC base pair, e.g. $k_{\kappa\chi} \approx k_{GC}$ and $k_{\kappa\chi}^{back} \approx k_{AT}^{back}$. In other words, the tautomers of $\kappa\chi$ are formed as slow as the tautomer of GC, but the backreaction to $\kappa\chi$ is as fast as in the case of AT. As a consequence there are fewer $\kappa\chi$ tautomers at a time than in the case of GC or AT.

Until here the protons in the DPTs were considered as classical mass points. In recent years experimental evidence has been aupplied that protons also possess a quantum nature, even at ambient conditions suitable for organisms. One surprising fact in support of the quantum nature is the thermal de Broglie wave length of a proton at 300 K, which is about 1 Å. This is one third of the distance between the protons forming the H-bond of the GC and AT base pair. Löwdin already pointed out, that the protons are not classical particles but "wave packets" obeying the laws of quantum theory, and thus, they may be subject to the well-known tunnel effect in the double well potentials of the DPTs. This effect implies that the genetic code cannot be completely stable, which furthermore means that this proton transfer over a distance of about 1 Å may be one of the driving forces in the evolution of living organisms [40, 95, 107]. For the determination of the tunnel (or reaction) rate the semi-classical method of Wentzel, Kramers and Brioullin (WKB) was used (for a review see ref. [108]). The WKB method alters the absolute values of the rate constants as determined by thermodynamics, but the aforementioned relations between the rate constants will remain qualitatively the same.

Now, let us focus on a situation where the quantum nature of protons may give a major qualitative difference. Chatzidimitriou-Dreismann [105] investigated some aspects of quantum entanglement between the protons in DNA base pairs. Based on the relative behaviour of the PESs of the DPTs it was argued that a possible dynamical correlation of the DPT reactions in the GC base pair may exist, since the PESs are different from each other, thus permitting the appearance of a Jordan-block structure in the restricted effective Hamiltonian of the DPT (see [105] for details). But in the $\kappa\chi$ base pair the existence of such a dynamical correlation is not possible, since, for symmetry reasons, the PESs of both DPTs are the same. Here, the involved PESs have been studied in considerable detail and

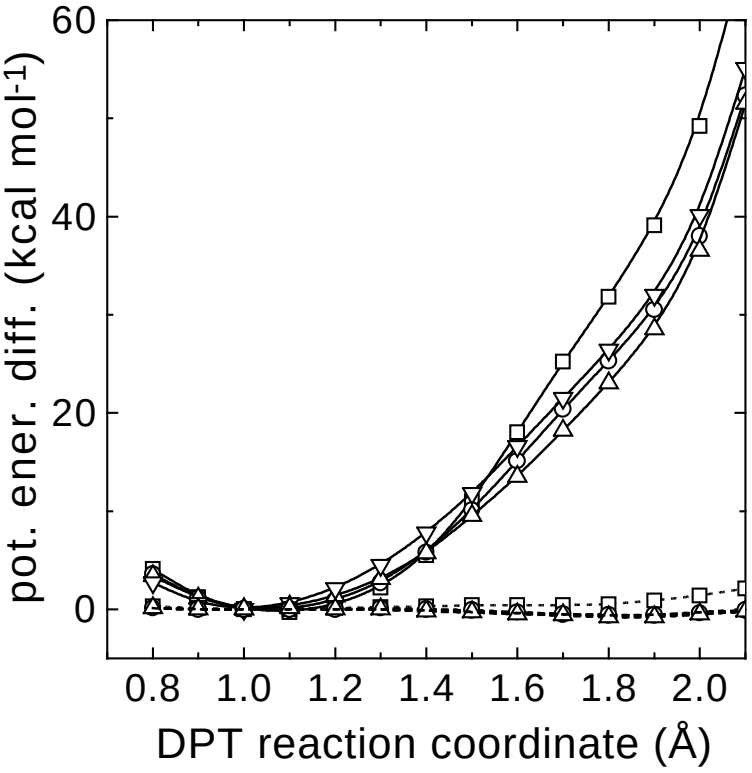

Figure 11.6: Shown are the differences of the PESs of the tautomer 2 and tautomer 1 of the GC base pair (solid line) as well as the $\kappa\chi$ base pair (dashed line) at different levels of theory: □: HF/STO-3G; ◯: HF/6-31G**; △: B3LYP/6-31G**; ▽: MP2/6-31G**. The solid lines are guides to the eye. Please note, that the difference for the $\kappa\chi$ base pair is constantly zero for the large basis sets.

indeed the PES of both DPT reactions in the $\kappa\chi$ base pair are identical, i.e. their difference vanishes (see fig. 11.6). Furthermore, the PESs of the DPT reactions in the GC base pair are different.

11.4 The phase stability argument

In the following, let us consider (some aspects of) the dynamics of the mentioned double proton transfer in the GC and $\kappa\chi$. In the light of the diabatic representation, the effective second order Hamiltonian matrix can be written as (see [109]):

$$H(R_1, R_2) = \begin{pmatrix} H_{11}(R_1) & C(R_1, R_2) \\ C(R_1, R_2) & H_{22}(R_2) \end{pmatrix} \tag{11.1}$$

As an example, the matrix elements H_{11} and H_{22} are considered to represent the matrix elements of the Hamiltonians of the investigated double proton transfers in the GC and $\kappa\chi$ base pair and the dynamical variable R_i represent the corresponding position of the two protons. H_{ii} equals the sum of the kinetic energy T_{ii} and the potential energy V_{ii} of the double proton transfer which have been studied in considerable detail. Of particular importance are the off-diagonal elements $C(R_1, R_2)$, which represent the quantum mechanical coupling of the two processes described by H_{11} and H_{22}, i.e. the double proton transfers of the GC or the $\kappa\chi$ base pair.

It is now important to realize that the coupling $C(R_1, R_2)$ cannot vanish for all positions of the moving protons for physical reasons. This has been made plausible in the following two ways [105]:

1. Adhering to quantum chemistry it has to be noted that a proton always drags a significant electron cloud with it, as it moves from one end of the H-bonds to the other. The coupling of the two protons is given by the repulsion (interaction) of the two electron clouds as the protons pass each other.

2. A more subtle reason for this coupling is given by the fact that protons are quantum objects. Therefore the protons exhibit an intrinsic delocalization, which can be estimated by the order of the de Broglie wavelength λ_{dB}. A short calculation of λ_{dB} of a (quasi-free) proton with a kinetic energy equal to the thermal energy k_BT (k_B: Boltzmann constant) gives $\lambda_{dB} \approx 1$Å. This numerical estimate of the protonic delocalization implies a partial overlap of the proton wave functions and therefore they can interfere, thus creating intrinsic quantum correlations between the protons.

Furthermore it was shown that the Hamiltonian matrix $H(R_1, R_2)$ is not diagonalizable under the condition

$$H_{11} = H_{22} \pm i2C \tag{11.2}$$

and shows the specific Jordan block structure [110]:

$$J(E) = \begin{pmatrix} E & 1 \\ 0 & E \end{pmatrix} \tag{11.3}$$

with $E = \frac{1}{2}(H_{11} + H_{22}) \pm \sqrt{(H_{11} - H_{22})^2 + 4C^2}$. This means that no similarity transformation is able to diagonalize H. Furthermore H has *only one* eigenvector [110] which has in the present case the explicit form [105]:

$$\Psi_j = \begin{pmatrix} 2C \\ H_{22} - H_{11} \end{pmatrix} \tag{11.4}$$

For more details see [105]. Of course, for a Hermitian matrix this can never happen.

11.5 Consequences for the GC and $\kappa\chi$ phase stability

In this section, the previous formal results and discussions are explicitly applied to the dynamics of the coupled double proton transfers (see section 11.1) in these base pairs. As discussed in the previous subsection, the diagonal elements H_{11} and H_{22} may be assumed to describe the two possible double proton transfers under consideration, and C may represent the quantum mechanical coupling (or 'interaction') between them. Furthermore, and in order to simplify the following remarks, let us restrict the considerations to the 'equidistant' (or 'cooperatively coupled') double proton transfers being defined by the requirement:

$$R_1 = R_2 \equiv R \tag{11.5}$$

By considering the pattern of the H-bonds in the base pairs (see figs. 10.1,10.2 and 10.3), the following point may be observed:
The triple H-bond pattern in $\kappa\chi$ (and $\kappa\pi$) exhibits a specific symmetry that is missing in GC.

Namely, the H-atom of the base contributing only one H-atom in the H-bond pattern (i.e., χ or π) is situated in the 'central' position, and the other two H-atoms 'belonging' to κ are situated symmetrically in the two 'outer' positions of the pattern. This symmetric distribution of the H-bonds of $\kappa\chi$ is clearly not present in the GC base pair.

This observation, however, appears to be crucial for the quantum dynamics of the protonic motions in the considered base pairs. Namely, the aforementioned 'symmetry' in $\kappa\chi$ also implies that the two double proton transfers (as discussed in section 11.1) are expected to have similar potential energy curves V_{ii}, for a significant part of the numerical range of R. Thus it can be assumed in first approximation:

$$H_{11}(R) = H_{22}(R) \quad \text{for } \kappa\chi \tag{11.6}$$

Remember that in the considered case of a complex symmetric effective Hamiltonian, the quantities H_{11} and H_{22} may be complex.

The same considerations suggest also that the corresponding 'asymmetry' of the H-bond pattern of GC gives rise to the following inequality (for a significant part of the range of R)

$$H_{11}(R) \neq H_{22}(R) \quad \text{for GC} \tag{11.7}$$

This result is in line with the different numerical values of the two potential barriers mentioned in section 11.2, where only the real parts of (the generally complex quantities) $H_{11}(R)$ and $H_{22}(R)$ were considered.

Some further reasoning leads to the result:

1. $\kappa\chi$: Equality 11.6 and the condition 11.2 can be fulfilled simultaneously only in the trivial case $C = 0$, which also makes the Hamiltonian H diagonal. But the latter point also implies that the two quantities H_{11} and H_{22} are uncoupled in the approximation under consideration. Thus it may be concluded that there do not exist either quantum correlations or phase stability of the aforementioned character between the two double proton transfers in $\kappa\chi$.

2. GC: Inequality 11.7 and the necessary condition 11.2 for the occurrence of Jordan blocks in the Hamiltonian H can be fulfilled simultaneously, for some value(s) of R, with non-vanishing coupling C. Thus, the quantum correlations between the two double proton transfers and the aforementioned increased phase stability (see section 11.4 above) may become effective in the GC base pair.

In order to illustrate these findings further, let us continue these investigations concerning the GC and $\kappa\chi$ base pair by taking into account the calculated potential energy surfaces of section 11.2. To be specific, let us consider in particular the following special case of the GC and $\kappa\chi$ dynamics:

$$H_{11}(R) \text{ and } H_{22}(R) : \text{ real} \tag{11.8}$$

$$C(R,R) \equiv i\overline{C}(R) \text{ with real } \overline{C}(R) \tag{11.9}$$

Due to the considered validity of the restriction $R_1 = R_2 \equiv R$ (see eq. 11.5) it may be assumed for the kinetic energies $T_{11} = T_{22}$. The condition 11.2 for the appearance of a Jordan block, for at least some 'position(s)' R_c, in the effective Hamiltonian matrix (by choosing the minus sign in it) reads now:

$$V_{11}(R_c) = V_{22}(R_c) + 2\overline{C}(R_c) \tag{11.10}$$

Here the dynamics of the $\kappa\chi$ base pair will be considered first. In the following the PES of the double proton transfer forming tautomer 1 of $\kappa\chi$ (see fig. 10.3) will be called V_{11} while V_{22} is associated with the double proton transfer forming tautomer 2 of $\kappa\chi$ (see fig. 10.3). The individual PES are depicted in figs. 11.4 and 11.5. The difference $V_{11} - V_{22}$ is given by the dashed lines in fig. 11.6 at the different levels of theory of section 11.2. As it is easily recognized, the difference of the PESs is for all more sophisticated levels of theory a constant at 0 kcal mol^{-1}, proving the PESs to be identical, i.e. $V_{11} = V_{22}$, and the quantum mechanical coupling to be absent, i.e. $C = 0$. This finding implies that the effective Hamiltonian of eq. 11.1 is already diagonal and no quantum correlation is present in the current approximation in the $\kappa\chi$ base pair.

Now let $V_{22}(R)$ represent the potential energy of the coupled transfer of the C-hydrogen with the 'central' G-hydrogen of the H-bond pattern (cf. tautomer

1 of GC fig. 10.2); $V_{11}(R)$ then represents the corresponding potential energy of the transfer of the two 'outer' protons in this bonding, (cf. tautomer 2 of GC fig. 10.2). For this special choice, the results of section 11.2 for the double-well potentials of these transfer processes are represented schematically in figs. 11.2 and 11.3. Their difference $V_{11} - V_{22}$ is depicted in fig. 11.6. These data reveal the following feature: As can be seen in fig. 11.6, the function $V_{11}(R) - V_{22}(R)$ is a monotonously increasing function in the 'reaction' coordinate R. This is by no means trivial, since this behaviour holds for (almost) all R-values between the two potential energy minima. Since the coupling C may fluctuate in the course of time (e.g. due to thermal disturbances), the possible value(s) of R_c may 'fluctuate', too.

Taking into account the above quasi-linearity of $V_{11}(R) - V_{22}(R)$ in the vincinty of the transition state and inspecting of the specific form of the single eigenvector $\Psi_J(R_c)$, (see eq. 11.4 in the case of a Jordan-block Hamiltonian) it is immediately concluded: The *direction* of $\Psi_J(R_c)$ is almost *independent* of the specific value of R_c, i.e.

$$\Psi_J(R_c) \sim \begin{pmatrix} i \\ 1 \end{pmatrix} \tag{11.11}$$

In other words, it follows that all the possible Jordan-block eigenvectors $\Psi_J(R_c)$ are related to (approximately) the same physical state, despite the fact that the effective Hamiltonian, eq. 11.1, is strongly R-dependent — an unexpected result indeed. Thus, under the condition that the numerical results of section 11.2 and ref. [111] are relevant, this striking finding physically means that the considered quantum mechanical phase stability (or rigidity) of the GC system is higher than the previous general derivations may have indicated.

At this stage, it is instructive to mention: the thermal motion (or: disturbance) may play a 'constructive role' in the stabilization of the DNA base pairs. This is certainly not expected, since usually it is assumed that the thermal motion disturbs (or even destroys) quantum effects. However, the situation here is more subtle. Namely, due to the thermal motion the distance between individual bases in base pairs does fluctuate in time. This implies a time dependence in the functional form and the strength of the quantities $H_{11}(R)$, $H_{22}(R)$ and $C(R, R)$. Therefore, condition 11.2 — being necessary for the occurrence of the specific quantum 'cooperativity' under consideration — can now be fulfilled more easily, during many time instants (or short time intervals), as time goes on. This qualitative consideration also illustrates the aforementioned dynamical character of the effect being revealed by theory, and it is conceptually in line with the recent work of *Prigogine* and coworkers concerning the 'constructive role' of irreversibility in self-organization processes (cf. [112]).

Here some concluding remarks may be appropriate. The present calculations show that for the determination of the PES of the DPT reactions it is necessary to

include the electron correlation energy in one or the other way to obtain reliable results. This holds especially for the transition states. At the correlated level both DPT reactions of the $\kappa\chi$ base pair are identical in energy. Furthermore, the $\kappa\chi$ base pair seems to have similar dynamics as compared with the AT and GC base pair, but the relative energy of the tautomers is considerably higher (about 7-9 kcal mol^{-1}).

In future work, it is intended to investigate the quantum entanglement and the associated dynamics of the DPT reaction based on the different PESs by the means of the CSM method [113–115], which determines the quasi-stationary energies and life times or tunnel rates of the eigenstates in the PES. Furthermore, an analysis of the dynamics of wave packets in these PES is planned.

12 Geometry-optimization

12.1 Introduction

Among the 29 possible base pairs the Watson-Crick base pairs guanine-cytosine (GC) and adenine-thymine (AT) have been most intensively investigated [92, 94, 96, 116–125], due to their occurrence in DNA. However, the size of these molecular systems restricted the application of *ab initio* quantum chemical models for the elucidation of the intrinsic properties of the PES of a DPT to relatively low levels of theory. To the best of my knowledge, two different approaches to attack this problem have been described in literature. The first one is to keep the coordinates of the heavy atoms fixed (e.g. at positions known from X-ray structural analysis of DNA), shifting the protons cooperatively along the H-bonds and scanning the potential energy surface as described in section 11. These calculations as well as their semi-empirical [120, 126, 127] counterparts, result in exceedingly high energy differences of the tautomers for either the AT as well as the GC base pair (cf. section 11.2). For example, the energy difference between the AT base pair and its tautomeric A^*T^* structure was predicted to be 36 kcal mol^{-1}, whereas for the GC base pair an energy difference of 34 and 57 kcal mol^{-1} (in the GC base pair two different DPTs are possible) was determined. However, in view of an at least 34 kcal mol^{-1} endothermic reaction it seems very unlikely that the tautomeric form of the base pairs is formed at all.

The second approach consists of identifying small, representative fractions of the entire base pair as benchmark systems, assuming that these smaller systems already account for most of the properties of the full base pair. Subsequently the benchmark system is fully geometry-optimized at a higher level of theory to obtain the minima and the transition state for the DPT. An example for this strategy is the computational study of the formamide dimer as a model for the AT pair by Hrouda et al. [128]. However, it is well known from many reactions in biology that the asymmetry involved is very important and should not be neglected [129], but the limitation of computer resources in the past forced the authors to choose a high symmetry model in order to reduce the computational effort.

Calculations employing full geometry optimization on the entire AT and GC base pair can be found in recent literature [130]. These calculations [130] throw a consistent light on the interaction energies and geometries only when including the electron correlation effects in one or the other way. But concerning the tautomeric structures of the AT and GC base pair no equivalents exist up till now. The most sophisticated calculations on the tautomeric structures are done in the HF picture with only the minimal basis sets like MINI-1 [131]. In the study, the energy difference between the tautomers dropped significantly down to about 10 kcal mol^{-1} and the transition structure for the double proton transfer is predict-

ed to be only marginally (0.2 kcal mol^{-1}) higher than the imino-enol tautomer structure.

In the light of these results it seems to be very unlikely that the system will remain in the tautomeric state even for a very short time [128], since RT (where R denotes the universal gas-constant) roughly equals 0.6 kcal mol^{-1} at room temperature (T = 298 K), and according to Boltzmann's law, vibrational states of the imino-enol tautomer near and above the transition state will be significantly populated.

In the following, the first consistent full geometry-optimizations of the complete AT base pair, its imino-enol tautomer and the transition state connected with the DPT are reported [92]. Furthermore the full geometry-optimization of the GC base pair, its tautomer and the transition state of the DPT reaction is presented at a level of theory adequate for obtaining predictions of (at least) semi-quantitative accuracy [92].

12.2 Computational methods

Standard quantum chemical calculations at the Hartree-Fock level, the valence electron correlated MP2 and the recently suggested DFT/HF hybrid (using the B3LYP functional as implemented in Gaussian 94 [104]) level of theory have been carried out employing a flexible, polarized 6-31G** one-particle basis set[5]. The geometries of all stationary points were optimized by analytical gradient techniques. The interaction energy of the hydrogen bonds in the dimers were corrected for basis set superposition errors (BSSE) using the standard counterpoise procedure. The effects of zero-point vibrational energies (ZPVE) were explicitly computed for the individual bases adenine, thymine, the corresponding imino and enol tautomers as well as the complexes by determining the harmonic frequencies from the analytically computed force constant matrices. All calculations utilized Gaussian94 [104] as installed on IBM RS/6000 workstations and CRAY computers.

12.3 Results and discussion

12.3.1 Benchmark system

One of the goals of the present study is to establish a computational method, which is capable of reproducing as accurately as possible the experimental structure and the binding energy of the AT base pair, which is experimentally known as 13 kcal mol^{-1} in the region of 298K [132]. To this end, a benchmark system consisting

[5]The smaller 3-21G basis set employed in initial calculations was soon discarded, since with this basis set the imino-enol tautomer could not be located regardless of the method used.

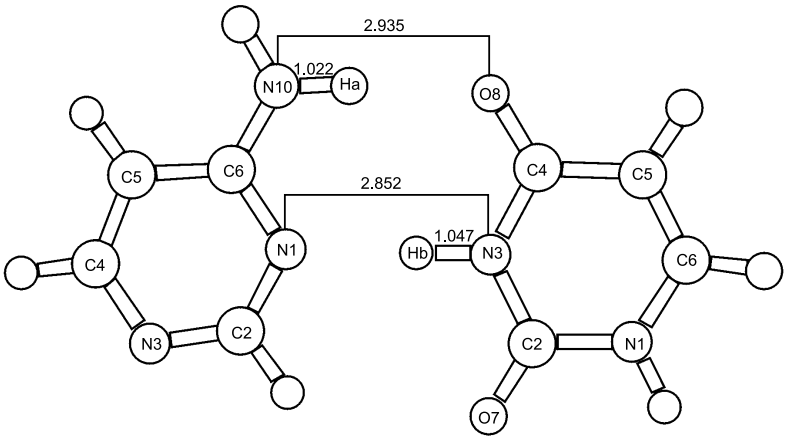

Figure 12.1: Shown is the structure of the benchmark system, consisting of the 6-membered ring of adenine (on the left) and the 6-membered ring of thymine (on the right). The atoms without label are hydrogen atoms. The given distances are taken of the B3LYP/6-31G** optimized structure and measured in Å.

of the main ring system of the base pair (fig. 12.1) has initially been optimized at various levels of theory. tabs. 12.1 and 12.2 contain a summary of the structure optimized at the HF/3-21G, HF/6-31G**, B3LYP/6-31G** and MP2/6-31G** levels of theory compared to the geometry of the whole AU base pair as obtained from X-ray single-crystal data [133].

Good agreement between the computationally predicted and experimental data can be observed for each of the 6-membered rings already at the HF/6-31G** level of theory. However, the H-bonding distances are significantly overestimated by about 0.1-0.2 Å at this level of approximation, underlining the well-known shortcomings of the HF approach for describing hydrogen bonds [134]. On the other hand, DFT/HF hybrid calculations, employing Becke's three parameter exchange functional with the Lee, Yang and Parr non-local correlation functional (B3LYP), which was applied successfully to intra-molecular H-bonds in a recent paper of Barone [135], show a very good agreement in intramolecular structure, but also that the error in the H-bonding distance is much smaller. The computed H-bonding distances in the benchmark system at the B3LYP/6-31G** level of theory differ by less than 0.05 Å from the X-ray data of the AU base pair (cf. tab. 12.1). The MP2/6-31G** level of theory does not further improve the structure as compared with the B3LYP/6-31G** level. Therefore, the calculations on the AT

Table 12.1 Bond lengths of the benchmark system as optimized by four different levels of theory ordered by the computational effort involved. Additionally, the corresponding bond lengths of the experimental AU base pair structure are given for comparison. The atom numbering is given in fig. 12.1.

method	experiment	HF	HF	B3LYP	MP2
basis set		3-21G	6-31G**	6-31G**	6-31G**
'adenine' ring					
N1-C2	1.35	1.333	1.323	1.338	1.326
C2-N3	1.34	1.318	1.310	1.330	1.317
N3-C4	1.36	1.347	1.335	1.350	1.337
C4-C5	1.36	1.365	1.366	1.380	1.369
C5-C6	1.41	1.409	1.408	1.416	1.406
C6-N1	1.36	1.338	1.331	1.353	1.338
C6-N10	1.36	1.339	1.340	1.349	1.340
'thymine'-ring					
N1-C2	1.36	1.382	1.375	1.399	1.381
C2-N3	1.38	1.364	1.367	1.380	1.367
N3-C4	1.37	1.374	1.378	1.393	1.378
C4-C5	1.42	1.455	1.460	1.457	1.451
C5-C6	1.34	1.329	1.330	1.351	1.338
C6-N1	1.36	1.373	1.369	1.372	1.365
C2-O7	1.22	1.216	1.195	1.220	1.206
C4-O8	1.27	1.226	1.203	1.234	1.218
H-bonds					
N10-Ha		1.008	0.999	1.022	1.010
N3-Hb		1.032	1.013	1.047	1.031
N1-Hb		1.754	1.990	1.805	1.828
O8-Ha		1.971	2.085	1.914	1.947
N10-O8	2.95	2.975	3.081	2.935	2.956
N1-N3	2.82	2.786	3.002	2.852	2.860

Table 12.2 Bond angles of the benchmark system as optimized by four different levels of theory ordered by the computational effort involved. Additionally, the corresponding bond angles of the experimental AU base pair structure are given for comparison. For the atom numbering see fig. 12.1

method basis set	experiment	HF 3-21G	HF 6-31G**	B3LYP 6-31G**	MP2 6-31G**
'adenine' ring					
N1-C2-N3	129.1	124.9	127.4	127.4	127.2
C2-N3-C4	109.5	116.3	114.9	114.6	114.8
N3-C4-C5	128.0	123.1	123.7	123.8	123.8
C4-C5-C6	117.4	117.4	116.5	117.1	116.8
C5-C6-N1	117.4	118.7	119.9	119.4	119.5
C6-N1-C2	118.5	119.7	117.5	117.7	117.9
C5-C6-N10	124.4	123.4	122.3	123.2	122.9
'thymine'-ring					
N1-C2-N3	116.4	113.8	114.2	113.6	113.9
C2-N3-C4	123.4	127.8	127.0	127.1	127.1
N3-C4-C5	117.7	114.6	114.9	115.0	115.0
C4-C5-C6	117.7	119.0	118.9	119.4	119.1
C5-C6-N1	123.3	122.2	122.0	121.5	121.7
C6-N1-C2	121.1	122.6	123.1	123.4	123.2
N1-C2-O7	123.7	122.4	122.1	121.9	122.0
N3-C4-O8	118.0	129.8	120.8	120.8	120.8
H-bonds					
C6-N10-Ha		120.0	120.3	120.4	120.3
C2-N3-Hb		115.4	115.9	115.7	115.8
C6-N1-Hb		125.1	125.4	124.4	124.3
C4-O8-Ha		123.9	125.7	124.0	124.1
N10-Ha-O8		173.9	174.8	176.5	176.0
N3-Hb-N1		178.3	178.1	179.3	179.7

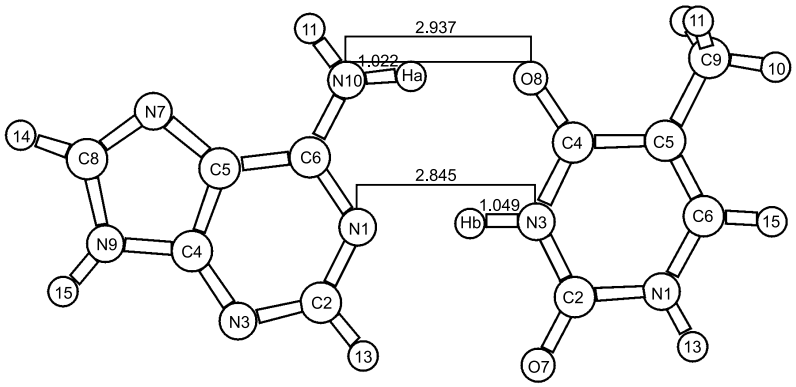

Figure 12.2: Depicted is the structure of the AT base pair as optimized by the B3LYP/6-31G** approach. The distances are measured in Å.

base pair are carried out at the B3LYP/6-31G** level.

12.3.2 AT base pair system

Geometry tabs. 12.3 and 12.4 summarize the optimized B3LYP/6-31G** geometry of the whole AT base pair (cf. fig. 12.2). Additionally the X-ray structure [133] is given for comparison. The agreement of the experimental and calculated structure is good for bond lengths and angles. Of course, the X-ray structure data have to be seen with an appropriate amount of caution, since the experimental data are not measured in a single AT base pair, but in an ApU duplex, in which the C9 methyl group is missing. The lengths of bonds, in which N3 of thymine or N9 of adenine participate, seem to be slightly overestimated (cf. tab. 12.3), which is probably caused by the replacement of the desoxyribose ring with a single hydrogen atom. Nevertheless the H-bonding distances between the bases should not be affected by this.

Here the improvements over previous calculations will be stressed. First the HF/MINI-1 [131] calculations generally overestimate the intramolecular bond lengths, whereas the H-bonding distance is drastically underestimated, which is reflected in the exceedingly high H-bonding energy of 17 kcal mol^{-1} as compared with 13 kcal mol^{-1} in the experiment [132]. The optimizations at the HF level with various split valence basis sets [124, 125] improve the description of the intramolecular structure, but do not succeed in predicting the H-bonds properly. The present approach improves the predicted structure mainly in the intermolecular separation. The H-bonding distances are computed as 2.845 Å for the symmetric (N-H N) and 2.937 Å for the asymmetric (N-H O) H-bond, respectively, which corresponds to an error below 1% (cf. tab. 12.3). A similar agreement can

Table 12.3 Bond lengths of the optimized B3LYP/6-31G** structure of the AT base pair, the imino-enol tautomer (A^*T^*) and the transition state of the DPT reaction (AT^{TS}). The bonds are defined by the participating atoms (cf. figs. 12.2 and 12.3)

	Bond	experiment Å	AT Å	A^*T^* Å	
adeninie	N1-C2	1.352	1.348	1.357	1.357
	C2-N3	1.346	1.332	1.312	1.312
	N3-C4	1.362	1.342	1.358	1.358
	C4-C5	1.359	1.398	1.393	1.394
	C5-C6	1.411	1.415	1.434	1.432
	C6-N1	1.361	1.353	1.390	1.388
	C5-N7	1.369	1.384	1.379	1.379
	N7-C8	1.329	1.311	1.312	1.312
	C8-N9	1.350	1.381	1.380	1.380
	N9-C4	1.370	1.377	1.373	1.373
	C6-N10	1.358	1.342	1.303	1.305
	N10-H11		1.007	1.016	1.016
	C8-H14		1.082	1.081	1.082
	N9-H15		1.009	1.009	1.009
	C2-H13		1.088	1.086	1.086
thymine	N1-C2	1.361	1.393	1.406	1.406
	C2-N3	1.377	1.380	1.371	1.370
	N3-C4	1.372	1.389	1.331	1.333
	C4-C5	1.420	1.465	1.452	1.453
	C5-C6	1.335	1.353	1.359	1.358
	C6-N1	1.358	1.376	1.367	1.367
	C2-O7	1.221	1.221	1.228	1.228
	C4-O8	1.279	1.236	1.299	1.295
	C5-C9		1.500	1.501	1.501
	C9-H10		1.093	1.094	1.094
	C9-H12		1.095	1.095	1.095
	N1-H13		1.009	1.010	1.010
	C6-H15		1.085	1.085	1.085
H-bonds	N10A-Ha		1.022	1.446	1.397
	O8T-Ha		1.918	1.102	1.128
	N3T-Hb		1.049	1.705	1.680
	N1A-Hb		1.796	1.056	1.061
	N10A-O8T	2.95	2.937	2.545	2.524
	N1A-N3T	2.82	2.845	2.754	2.735

Table 12.4 Bond angles as optimized by the B3LYP/6-31G** approach of the AT base pair, the imino-enol tautomer (A^*T^*) and the transition state of the DPT reaction AT^{TS}. The bond angles are defined by the participating atoms (cf. figs. 12.2, and 12.3)

	angle	experiment degrees	AT degrees	A^*T^* degrees	
adenine	N1-C2-N3	129.1	128.2	125.6	125.7
	C2-N3-C4	109.5	111.2	111.4	111.4
	N3-C4-C5	128.0	127.2	128.2	128.2
	C4-C5-C6	117.4	116.4	118.3	118.2
	C5-C6-N1	117.4	117.3	111.5	111.7
	C6-N1-C2	118.5	119.7	125.0	124.9
	C4-C5-N7	111.1	111.5	111.4	111.4
	C5-N7-C8	103.9	104.0	104.2	104.2
	N7-C8-N9	112.6	113.3	112.9	112.9
	C8-N9-C4	106.3	106.7	106.8	106.8
	N9-C4-C5	106.0	104.4	104.7	104.6
	C5-C6-N10	124.4	123.2	129.2	129.0
	C6-N10-H11		118.7	111.3	111.6
	N3-C2-H13		117.2	120.1	120.1
	N7-C8-H14		125.2	125.4	125.4
	C8-N9-H15		127.5	127.6	127.5
thymine	N1-C2-N3	116.4	113.3	115.6	115.6
	C2-N3-C4	123.6	127.0	122.4	122.6
	N3-C4-C5	117.7	116.1	122.2	121.9
	C4-C5-C6	117.7	117.7	115.4	115.5
	C5-C6-N1	123.3	122.3	121.3	121.3
	C6-N1-C2	121.1	123.6	123.0	123.0
	N1-C2-O7	123.7	122.3	120.3	120.3
	N3-C4-O8	118.0	120.9	120.4	120.4
	C4-C5-C9		118.3	120.6	120.5
	C5-C9-H10		111.3	111.1	111.1
	C5-C9-H12		110.9	111.1	111.1
	C6-N1-H13		121.3	121.4	121.4
	C5-C6-H15		122.3	122.5	122.5
H-bond	C6A-N10A-Ha		120.6	127.6	127.2
	C4T-O8T-Ha		124.5	114.4	114.9
	C2T-N3T-Hb		115.7	111.6	111.9
	C6A-C1A-Hb		122.8	115.5	115.5
	N10A-Ha-O8T		174.4	175.2	175.2
	N3T-Hb-N1A		179.9	172.0	172.4

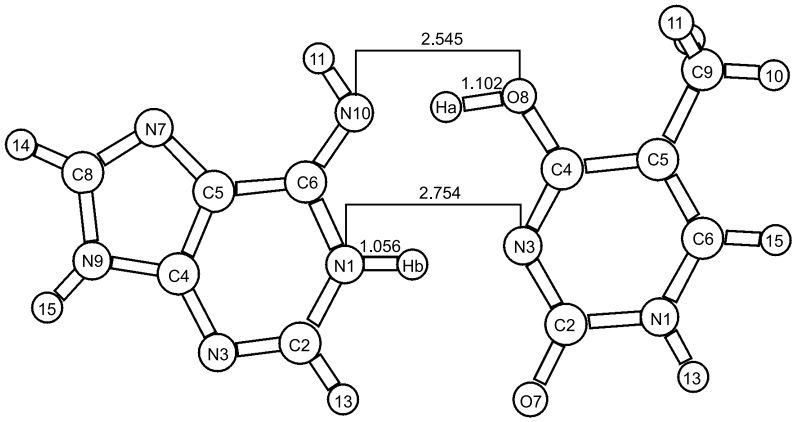

Figure 12.3: Shown is the structure of the imino-enol tautomer (A^*T^*) optimized at the B3LYP/6-31G** level of theory. Most impressive are the very short N10-O8 H-bond as well as the very long O8-Ha bond. The bond lengths are given in Å.

be achieved for the intramolecular bonding distances. In addition, the predicted bond angles are in good agreement with the experimental data, but are slightly better for the adenine part of the base pair than for the thymine part.

Furthermore, an isolated adenine has a pyramidal amino group, which is favoured marginally (0.01 kcal mol^{-1}) over a planar arrangement at the B3LYP/6-31G** level. In contrast, in the AT base pair the planar arrangement is favoured, i.e there is no negative eigenvalue of the associated Hessian matrix (cf. later in this section). Although it is not possible to decide on the planarity of the amino group of the adenine at this level of theory, the group will be very floppy anyway.

After assessing the accuracy that can be expected from the current B3LYP/6-31G** approach, the focus is now turned to the principal target of the present investigation, the imino-enol tautomer of the AT base pair and the transition state for the double proton transfer. The geometries are summarized in tabs. 12.3 and 12.4 as well as in fig. 12.3. At first sight the structures are very similar to those obtained from previous calculations. There is a significant shortening of the H-bonding distance in the asymmetric H-bond of the imino-enol tautomer of almost 0.4 Å in contrast to the earlier predictions. Florian et al. [131] found a much smaller shortening for the asymmetric H-bond at HF/MINI-1 level (about 0.2 Å). Also noteworthy is the very long O-H bond in the asymmetric H-bond of 1.102 Å. The length of a normal O-H bond is slightly less than 1 Å. In contrast, the N1-Hb bond is predicted to be 1.056 Å long, similar to the bond length of the N3-Hb bond in the AT base pair. The symmetric H-bond in the present calculation is shortened by 0.09 Å, in line with the HF/MINI-1 level of theory (0.07 Å) [131]. Experimental indications for such a shortening during a tautomerization had been

Table 12.5 The absolute energy E, the basis set superposition error BSSE, the zero-point vibrational energy ZVPE and the absolute entropy S^{298} at 298 K are given as calculated by the B3LYP/6-31G** approach for the isolated bases (A,T), their tautomers (A^*,T^*), the corresponding Watson-Crick base pairs (AT, A^*T^*) and the associated transition state of the DPT (AT^{TS}). All values for AT^{TS} are estimated except for the absolute energy.

molecule	absolute energy H	BSSE kcal mol^{-1}	ZPVE kcal mol^{-1}	S^{298} cal mol^{-1} K^{-1}
A	-467.3313907		70.1	82.2
A^*	-467.3121468		70.9	82.8
T	-454.1488137		72.2	86.9
T^*	-454.1279746		72.0	86.4
AT	-921.5063657	4.1	143.7	133.6
AT^{TS}	-921.4807176	4.2	141.9	
A^*T^*	-921.4853641	4.2	141.9	131.1

found by different authors [136, 137]. A question, which cannot be answered yet, is whether the phosphate-backbone of the DNA does allow such a torsion in the plane of the base pair or not, as a consequence of the dramatic shortening of the asymmetric H-bond.

Besides the changes in the H-bonds there are smaller changes in the neighbouring bonds. The bond C6-N10 of the adenine part is shortened from 1.342 Å to 1.303 Å indicating a double bond character. Furthermore, the C4-O10 bond is elongated from 1.236 Å to 1.299 Å as a sign of the change from a double to a single bond. The adjacent and more distant bonds are not significantly influenced. Especially the bonds of the ring systems in the AT base pair and its A^*T^* tautomer do not show the bond lengths of localized single and double bonds (e.g. a typical C-C single bond is about 1.5 Å long, cf. the C5-C9 bond of thymine), but intermediate ones of about 1.35 Å. This indicates that the aromatic systems survive the tautomerization reaction.

Now focus is moved to the transition state structure AT^{TS} of the DPT in the AT base pair. First it has to be mentioned that due to a bug in the program package used it was not possible to identify the true stationary point. When approaching the transition state structure the whole base pair system started drifting in real space and periodically diverging from and again converging to the stationary point, but never reaching it. Here is reported the structure with the smallest forces acting on the nuclei after the self consistent field process for the electrons has converged. The maximal force is about twice as large as the convergence criterion and acts on Hb. All other forces have reached the criterion and also the root mean

Table 12.6 Given are the relative energy ΔE^{B3LYP}, the difference ΔBSSE, the relative zero-point vibrational energy ΔZPVE, ΔE^0 the relative energy corrected for ΔZPVE and ΔBSSE and the relative entropies ΔS^{298} of the association reactions and the tautomerization. For each quantity the difference of the sum of the products and the sum of the educts are given.

reaction	ΔE^{B3LYP} kcal mol^{-1}	ΔBSSE kcal mol^{-1}	ΔZPVE kcal mol^{-1}	ΔE^0 kcal mol^{-1}	ΔS^{298} cal mol^{-1} K^{-1}
A+T $\rightarrow$ AT	-16.4	4.1	1.4	-10.9	-35.6
A*+T* $\rightarrow$ A*T*	-28.4	4.2	-0.9	-25.1	-38.1
AT $\rightarrow$ A*T*	13.2	0.1	-1.8	11.5	-2.5
AT $\rightarrow$ AT^{TS}	16.2	0.1	-1.8	14.5	
A*T* $\rightarrow$ AT^{TS}	3.0	0.0	0.0	3.0	

square of the forces has converged. Furthermore, the estimated Hessian matrix shows only one negative eigenvalue, suggesting the structure to be the transition state.

The structure of the proposed transition state AT^{TS} is very similar to the A*T* tautomer structure. The differences are almost neglegible except for the degrees of freedom, which describe the H-bonding system and its nearest neighbours. I.e. also in the AT^{TS} the aromatic system survives. The O8T-Ha bond of the asymmetric H-bond is elongated further by 0.026 Å to 1.128 Å, whereas the N1A-Hb bond of the symetric H-bond is only marginally stretched by 0.005 Å. Remember, the strongest remaining force acted on Hb. Probably the N1A-Hb bond is a little longer. But most interesting the adenine and the thymine part aproach each other in the AT^{TS} structure still a bit further, leaving a very short asymetric H-bond, N10A-O8T, with 2.524 Å and the symetric one, N1A-N3T, with 2.735 Å.

Vibrational Spectra All stationary points are optimized in the C_S point group symmetry, i.e. the molecules are constrained to the planarity of the two bases forming the pair. In order to identify the located stationary points as minima and saddle point, respectively, and to obtain the contributions from zero-point vibrational energies (ZPVE), the eigenvalues of the force constant matrix and the harmonic frequencies have been determined. The force constant matrix for the individual bases, the imino and enol tautomer as well as their complexes (AT and A*T*) have been determined, proving them to be true minima with a positive definite Hessian matrix. Inspecting the harmonic frequencies of the AT base pair and its A*T* tautomer it is evident, that there are several harmonic frequencies in the

A^*T^* tautomer (e.g. at 1799 cm^{-1}, 1779 cm^{-1}, 1685 cm^{-1} or 1645 cm^{-1}) which are associated with vibrations in which both protons of the H-bonds participate significantly at the same time. In contrast, the vibrations of the AT base pair are always dominated by one of both protons. There has to be some coupling of the vibrations to enter the DPT reaction; a possible explanation has been given in section 11.4. Prospective vibrations for this coupling are located at 3408 cm^{-1} and 2952 cm^{-1}. Furthermore, the harmonic frequency of the vibration of the symmetric H-bond is shifted to lower energies in the A^*T^* tautomer, i.e. from 2952 cm^{-1} to 2813 cm^{-1}, whereas the vibration of the asymmetric H-bond, which has a harmonic frequency of 3408 cm^{-1} in the AT base pair, vanishes in the A^*T^* tautomer. Altogether it seems to be easier to start the DPT reaction from the A^*T^* tautomer, if no quantum mechanical coupling of the kind described in section 11.4 is assumed.

The results of the zero-point vibrational energies are summarized in tab. 12.5. Here the relevant differences ΔZPVE for the association reaction A+T $\longrightarrow$ AT (ΔZPVE(AT))[6] and A^*+T^* $\longrightarrow$ A^*T^* (ΔZPVE(A^*T^*)) of the complexes as well as the tautomerization reaction AT $\longrightarrow$ A^*T^* (ΔZPVE(taut)) are addressed. The AT base pair has a slightly higher ZPVE than the individual bases alone, therefore ΔZPVE(AT) equals 1.4 kcal mol^{-1}. In the case of the formation of the imino-enol tautomer out of the imino and enol bases it is the other way round, therefore ΔZPVE(A^*T^*) equals -0.9 kcal mol^{-1}, indicating the ZPVE of the imino-enol complex to be smaller than the sum of the ZPVEs of the bases. The small ZPVE of the imino-enol tautomer is further expressed by a comparison with the ZPVE of the AT base pair which is by 1.8 kcal mol^{-1} smaller than in the latter, denoted by ΔZPVE(taut)=-1.8 kcal mol^{-1}. The ZPVE of the transition state structure is only estimated, since no real stationary point has been found. But, since the structure of AT^{ts} is almost identical to the structure of the tautomer A^*T^* it is reasonable to estimate the ZPVE's of both structures to be identical. Furthermore, the impact of the ZPVE for the prediction of tautomeric equilibria is well-known [116] and the results obtained here can be compared with these predictions.

The recent study of Santamaria and Vaszques [125] revealed that for a good approximation the adenine and thymine base in the AT base pair can be treated as independent oscillators. This conclusion is based on the small anharmonicity they found for the AT base pair at the local spin density functional level of theory. Hence, the estimated ZPVE of the AT structure as well as that of the imino-enol structure should be the sum of the zero-point vibrational energies of the individual bases adenine, thymine, adenine*, and thymine* (cf. tab. 12.5). In this approximation, the difference between the ZPVEs between the AT and A^*T^* pair amounted

[6]The arrow $\longrightarrow$ represents in the present context not an irreversible reaction, but an equilibrium.

to 0.6 kcal mol^{-1}. Compared to the actual value of ΔZPVE(taut) this is by 1.2 kcal mol^{-1} too small, only one third of the actual value. ΔZPVE(taut) may be small compared to the total ZPVE of the AT base pair (144 kcal mol^{-1}) or its A*T* tautomer (142 kcal mol^{-1}), but compared to the tautomerization energy of 11.5 kcal mol^{-1} it is a significant amount of more than 10 %. Therefore, it is necessary to take the influence of the ZPVE into account explicitly.

For later discussion the absolute entropies of the isolated bases and the base pairs are given in tab. 12.5. During the association reaction A+T $\longrightarrow$ AT and A*+T* $\longrightarrow$ A*T*, the number of molecules is reduced by one. Therefore a reduction (or destruction) of entropy during these reactions is expected, since the entropy is connected to the number of particles in statistical mechanics by $S = k(N\ln(N) - \sum_i n_i \ln(n_i))$, with k the Boltsmann constant, N the total number of particles and n_i the number of particles in the ith state. Calculating the corresponding entropy differences ΔS^{298} using the absolute entropies S^{298} of tab. 12.5 one finds ΔS^{298}(AT)=-36 cal mol^{-1} K^{-1} and ΔS^{298}(A*T*)=-38 cal mol^{-1} K^{-1}. The tautomerization reaction shows also a negative entropy difference with ΔS^{298}(taut)=-2.5 cal mol^{-1} K^{-1}; cf. tab. 12.6.

Energies and enthalpies The interaction energy for the AT and A*T* base pair was computed according to

$$\Delta E^0 = E(XY) - E(X) - E(Y) + \Delta ZPVE(XY) + \Delta BSSE(XY) \quad (12.1)$$

with

$$\Delta ZPVE(XY) = ZPVE(XY) - ZPVE(X) - ZPVE(Y) \quad (12.2)$$

where E(XY) is the energy of the base pair, ZPVE(XY) its zero-point vibrational energy; E(X) denotes the energy of the individual base X optimized in its own basis and ZPVE(X) the corresponding zero-point vibrational energy. BSSE(XY) is the basis set superposition error computed in the standard Boys-Bernardi counterpoise scheme [138]. For a summary of the energies of the isolated bases and the base pairs, as well as the zero-point vibrational energies and the BSSE see tabs. 12.5 and 12.6. If the ZVPE and BSSE are not taken into account, the B3LYP calculations predict the strength of the H-bonding as 16 kcal mol^{-1}, overestimating the experimentally known H-bonding energy in the AT base pair of 13 kcal mol^{-1} [132]. The BSSE for this particular system has been determined a couple of times at the HF-SCF level with different basis sets [124, 125, 131]. For the 6-31G** basis set, which is used in all previous calculations, in literature exists only an estimate of the BSSE based on an HF/3-21G optimized structure which results in 2.4 kcal mol^{-1} [124]. For the fully optimized structure and employing the B3LYP level of theory instead of the HF level of theory a slightly higher value for the

BSSE with this basis set of 4.1 kcal mol^{-1} is obtained. Correcting the interaction energy for the BSSE and zero-point vibrational energy effects ΔZPVE(AT) leads to a predicted strength of the two hydrogen-bonds of 11 kcal mol^{-1}, which is in good agreement with the experimental figure of 13 kcal mol^{-1} [132]. It should, however, be kept in mind that the computed H-bond strength refers to 0K, whereas the experiment [132] was carried out at 300K. But one can expect an even better agreement between the two results after a temperature correction is being added, since the temperature dependence of the interaction enthalpy, according to ref. [139], is always negative, i.e. the energy release during the formation of the super molecule increases with increasing temperature.

The energy release of the system accompanied by the formation of the imino-enol tautomer from a tautomerized adenine and thymine, i.e. for the reaction A*+T* $\longrightarrow$ A*T*, equals 25 kcal mol^{-1}. This is more than twice as large as the energy release by the formation of the AT base pair from isolated adenine and thymine molecules. Most of this significantly larger interaction energy for the imino-enol tautomer is due to the fact that the separated molecules of tautomerized adenine and thymine are 26 kcal mol^{-1} less stable than isolated adenine and thymine molecules. On the other hand, the shortening of the H-bond distances already discussed above indicate a higher interaction energy ΔE for the imino-enol tautomer than compared to the conventional AT base pair.

In tab. 12.6 a summary of the energetical aspects of the formation of the AT base pair and its tautomer from the corresponding isolated molecules and of the barrier height of the tautomerization is given. As expected, the AT structure is significantly more stable than the tautomeric structure by more than 11 kcal mol^{-1}. It is interesting to note, that the energy release of the AT base pair formation is in this approximation the same as the energy difference of the AT base pair and the A*T* tautomer. The activation barrier connected with the transition state of the DPT is predicted as 3 kcal mol^{-1} with respect to the A*T* pair. This is much less than predicted by calculation in the 'frozen core' approximation (cf. section 11.2). This is possibly due to the fact that the optimized transition state structure AT^{TS} shows a significantly smaller distance between the adenine and thymine part. During the optimization process it has been observed that the relative energy of the transition state structure with respect to the A*T* structure is reduced as the two subsystems adenine and thymine approach each other. On the other hand, earlier calculations with complete geometry optimization at the HF/MINI-1 level of theory resulted [131] in an energy difference between the two forms of the base pair of about 10 kcal mol^{-1}. These calculations also predicted an almost negligible activation barrier of 0.2 kcal mol^{-1} with respect to the A*T* pair for the transition state structure of the DPT. Thus, these calculations disagree significantly with respect to the barrier height of the DPT, which is of prime importance for the mean lifetime of the tautomeric state of the double proton transfer. One origin

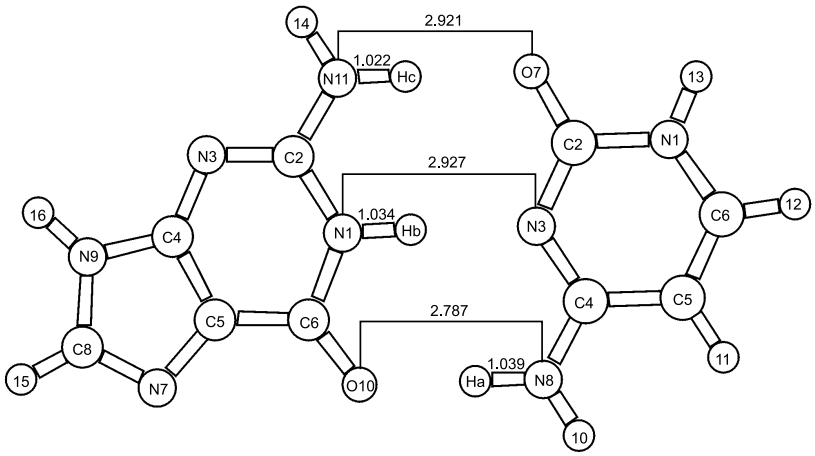

Figure 12.4: Depicted is the structure of the GC base pair as optimized by the B3LYP/6-31G** approach. The distances are measured in Å.

of this difference might be the possibly overestimated strength of the H-bonds in the earlier calculations, which is reflected in too short an intermolecular distance of 2.75 Å in the AT base pair. At the same time, the transition state is energetically the more lowered as the intermolecular distance is reduced in the complex. The present calculations predict 2.95 Å as the H-bonding distance of the central H-bond in the AT base pair being in very good agreement with the X-ray structures (see tab. 12.3) and, consequently, a more substantial barrier for the double proton transfer.

12.3.3 GC base pair system

Geometries The structure of the GC base pair optimized at the B3LYP/6-31G** level of theory is presented in fig. 12.4. The bond lengths are given in tab. 12.7 as well as the bond angles are given in tab. 12.8. Additionally, the experimental X-ray structure [140] is given in the tables for comparison.

The predicted structure of the GC base pair is of similar quality as in the case of the AT base pair compared with the experiment. However, in the experiment a GpC duplex stabilized by cations has been investigated. Therefore, an appropriate amount of caution has to be applied in a direct comparison of the calculated and experimental structure. Just as in the case of the AT base pair there are some systematic deviations which are due to the approximations applied in the calculation. Most evident is the overestimated bond length of the bond connecting the two rings of the purine which is predicted to be 1.400 Å instead of 1.358 Å (see also bond C4-C5 of adenine in the AT base pair). Furthermore, there are some problems connected with the carbonyl system of the guanine base, which is ex-

Table 12.7 Bond lengths of the optimized B3LYP/6-31G** structure of the GC base pair, the imino-enol tautomer (G^*C^*) and the transition state of the DPT reaction (GC^{TS}). The bonds are defined by the participating atoms (cf. figs. 12.4, 12.5 and 12.6)

	bond	experiment Å	GC Å	G^*C^* Å	GC^{TS} Å
guanine	N1-C2	1.382	1.376	1.368	1.380
	C2-N3	1.362	1.327	1.344	1.339
	N3-C4	1.350	1.349	1.334	1.336
	C4-C5	1.358	1.400	1.402	1.399
	C5-C6	1.452	1.431	1.410	1.412
	C6-N1	1.390	1.408	1.346	1.369
	C5-N7	1.397	1.385	1.389	1.388
	N7-C8	1.320	1.306	1.305	1.305
	C8-N9	1.361	1.387	1.389	1.390
	N9-C4	1.379	1.372	1.374	1.373
	C6-O10	1.206	1.241	1.316	1.291
	C2-N11	1.331	1.350	1.355	1.347
	N11-H14		1.006	1.005	1.006
	C8-H15		1.082	1.082	1.082
	N9-H16		1.009	1.008	1.008
cytosine	N1-C2	1.394	1.411	1.394	1.400
	C2-N3	1.345	1.359	1.368	1.362
	N3-C4	1.318	1.339	1.394	1.380
	C4-C5	1.445	1.445	1.455	1.454
	C5-C6	1.341	1.356	1.349	1.349
	C6-N1	1.357	1.361	1.374	1.367
	C2-O7	1.236	1.235	1.227	1.235
	C4-N8	1.327	1.336	1.291	1.303
	N8-H10		1.007	1.017	1.014
	C5-H11		1.082	1.082	1.082
	C6-H12		1.084	1.084	1.084
	N1-H13		1.010	1.009	1.009
H-bonds	N8-Ha		1.039	1.690	1.400
	O10-Ha		1.748	1.011	1.102
	N1-Hb		1.034	1.830	1.327
	N3-Hb		1.893	1.052	1.294
	N11-Hc		1.022	1.014	1.022
	O7-Hc		1.899	1.964	1.784
	N8-O10	2.91	2.787	2.695	2.501
	N3-N1	2.95	2.927	2.881	2.621
	O7-N11	2.86	2.921	2.979	2.805

Table 12.8 Bond angles as optimized by the B3LYP/6-31G** approach of the GC base pair, the imino-enol tautomer (G^*C^*) and the transition state of the DPT reaction GC^{TS}. The bond angles are defined by the participating atoms (cf. figs. 12.4, 12.5 and 12.6)

	angle	experiment Å	GC Å	G^*C^* Å	GC^{TS} Å
guanine	N1-C2-N3	122.0	123.4	126.8	125.7
	C2-N3-C4	112.2	122.3	112.0	112.5
	N3-C4-C5	129.7	129.3	127.7	127.8
	C4-C5-C6	118.7	117.9	115.4	116.5
	C5-C6-N1	111.1	111.3	119.0	117.0
	C6-N1-C2	126.3	125.9	119.2	120.4
	C4-C5-N7	111.9	111.1	111.2	111.3
	C5-N7-C8	102.5	104.4	104.1	104.1
	N7-C8-N9	114.0	113.0	113.4	113.3
	C8-N9-C4	106.0	106.8	106.6	106.6
	N9-C4-C5	105.6	104.6	104.6	104.7
	N1-C6-O10	120.7	120.1	121.0	121.6
	N1-C2-N11	118.4	116.6	116.6	117.1
	C2-N11-H14		116.6	117.1	116.1
	N7-C8-H15		125.6	125.4	125.4
	C8-N9-H16		127.8	127.7	127.7
cytosine	N1-C2-N3	118.8	117.2	114.7	116.9
	C2-N3-C4	121.8	121.2	126.3	122.6
	N3-C4-C5	120.5	121.7	115.4	118.5
	C4-C5-C6	117.0	117.0	119.4	118.8
	C5-C6-N1	121.7	120.2	121.2	120.2
	C6-N1-C2	120.1	122.6	123.0	123.0
	N3-C2-O7	122.7	124.7	124.5	125.0
	N3-C4-N8	120.9	117.7	117.3	117.8
	C4-N8-H10		120.0	112.0	113.6
	C4-C5-H11		121.6	119.5	119.9
	C5-C6-H12		123.1	122.9	123.4
	C6-N1-H13		121.8	121.8	122.1
H-bonds	C4-N8-Ha		120.7	125.9	124.3
	C6-O10-Ha		125.7	113.6	117.1
	C6-N1-Hb		115.4	121.1	119.6
	C4-N3-Hb		122.1	116.7	117.8
	C2-N1G-Hc		123.1	122.7	123.3
	C2-O7-Hc		118.5	117.3	116.2
	N8-Ha-O10		179.6	172.3	176.2
	N3-Hb-N1		177.9	176.8	177.9
	O7-Hc-N11		179.6	178.8	176.7

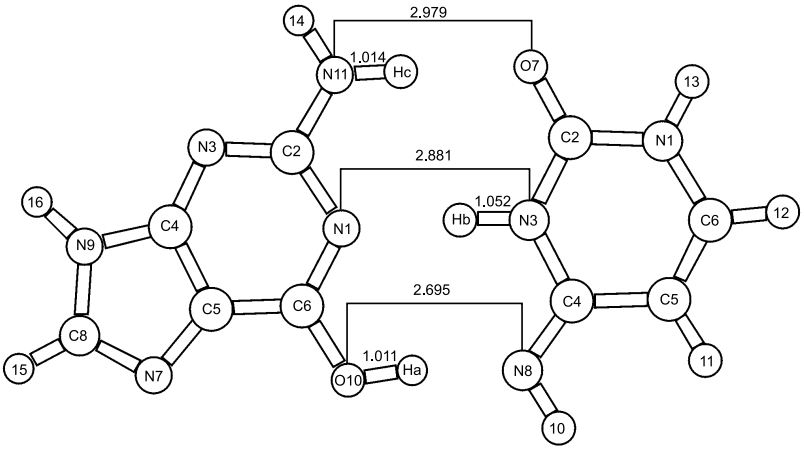

Figure 12.5: Shown is the structure of the imino-enol tautomer (G^*C^*) optimized at the B3LYP/6-31G** level of theory. The bond lengths are given in Å.

pressed by the overestimated bond length C6-O10 (1.241 Å instead of 1.206 Å; cf. tab. 12.7). Additionally, the electron density seems to be transfered to the neighbouring bonds, C5-C6, and, more important, to the H-bond O10-N8, which are therefore shortened. The mean deviation of the bond lengths in the predicted structure from the experimental ones [140] is 1.4% for the guanine part of the base pair. A by far better situation exists for the cytosine part with a mean deviation of 0.7% of the bond lengths. More pleasing is the situation that arises from the bond angles which are in very good agreement with the experiment [140], with a mean deviation of 0.9% and 0.4% for the guanine and cytosine part, respectively.

Concerning the H-bonds of the GC base pair the B3LPY/6-31G** approach predicts the central H-bond to be 2.927 Å which is in very good agreement with the experiment (2.95 Å, cf. ref. [140]). Even the H-bond facing the minor grove shows with a bond length of 2.921 Å a fair agreement with the experiment (2.87 Å), but as mentioned earlier there are some problems with the H-bond facing the major grove, which is predicted to be 2.787 Å instead of 2.91 Å. This is by 0.12 Å too short and causes, as will be seen later, too high an interaction energy.

Now, focusing on the H-bonds of the G^*C^* tautomer structure (see fig. 12.5 and tabs. 12.7 and 12.8) a shortening of the H-bonds participating in the DPT reaction is found, i.e. the bonds N8-O10 and N3-N1. This shortening of both H-bonds is less pronounced than in the tautomer of the AT base pair and none of the H-bonds is shortened by more than 0.1 Å. The largest shortening of 0.092 Å shows the N8-O10 H-bond, whereas the central H-bond N3-N1 is shortened by only half the amount (0.046 Å). Furthermore, the O7-N11 H-bond is elongated by 0.058 Å. This behaviour of all three H-bonds leads to a tilt of the bases against each other, just like in the A^*T^*, but less pronounced. In view of the shortening of

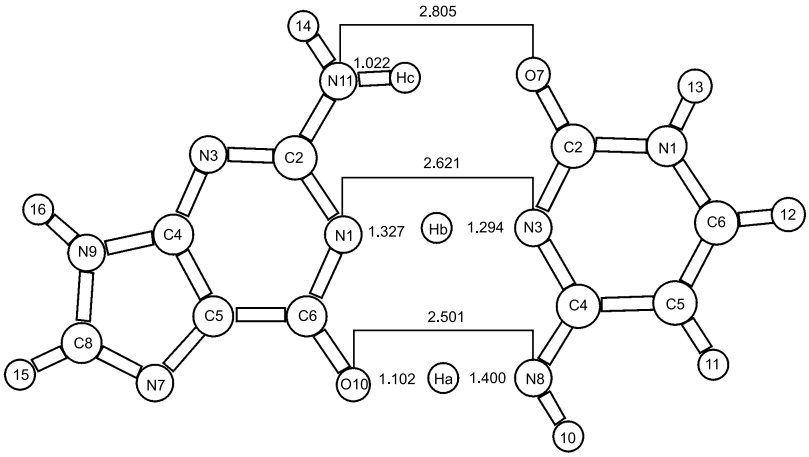

Figure 12.6: The transition state structure of the DPT GC→G*C* as optimized at the B3LYP/6-31G** level of theory. The bond length are given in Å.

the H-bonds involved in the DPT reaction and the elongation of the third H-bond, one may expect the H-bond interaction energy to remain about the same in the G*C* tautomer as in the GC base pair.

In the bonds adjacent to the H-bonds also some changes can be found. In the guanine part the C6-N1 bond is shortened from 1.408 Å to 1.346 Å and the C6-O10 bond is elongated from 1.241 Å to 1.316 Å indicating a change from a double bond to a single bond. Some changes are also evident in the cytosine part where the C4-N8 distance is shortened form 1.336 Å to 1.291 Å and the N3-C4 bond is elongated from 1.339 Å to 1.394 Å. In the rest of the ring systems there is only little change and the largest shortening or elongation is 0.017 Å. Therefore, the aromatic systems of the GC base pair seem to survive the DPT reaction and a large energy change is not to be expected.

Inspecting the transition state structure (see fig. 12.6 as well as tabs. 12.7 and 12.8) of the DPT reaction, GC → G*C*, an even more pronounced shortening of the H-bonds is easily recognized as compared with the G*C* tautomer. Furthermore, all three H-bonds are shortened in the transition state structure, not only in the H-bonds participating in the DPT reaction as in the G*C* tautomer structure. But it is still possible to distinguish between the different H-bonds. The bonds N8-O10 and N3-N1 are compressed by 0.28 Å and 0.30 Å respectively, which is about 10 % of their bond length, whereas the bond O7-N11 is only shortened by 0.12 Å which corresponds to about 4% of the bond length. The proton of the H-bond N3-N1 takes a position in the middle of the bond in the transition state structure, i.e. the bonds N3-Hb and N1-Hb are 1.294 Å and 1.327 Å long, respectively. In contrast, already in the asymmetric H-bond N8-O10 the proton is significantly shifted to the more electronegative O10 atom. The bond O10-Ha is

1.102 Å long, whereas the N8-Ha bond with 1.400 Å is significantly longer. In the third H-bond, which is not involved in the DPT reaction, only the distance of the heavy atoms, O7 and N11, is changed. The bond length N11-Hc remains unchanged at 1.022 Å.

Concerning the bonds of the ring systems, intermediate bond lengths are found between those in the GC base pair and its tautomeric G^*C^* structure. Even in the transition state the aromaticity of the ring systems seem to be retained.

Vibrational spectra Inspecting the vibrational spectra of the GC base pair, the G^*C^* tautomer and the transition state structure GC^{TS}, prooves the GC base pair to be a true minimum with a positive definite Hessian matrix. In the spectrum of the transition state structure a vibration with a negative harmonic frequency of -1248 cm^{-1} exists, which is connected with the protons Ha and Hb (cf. fig.12.6), identifying the associated vibrational mode as the driving force of the DPT reaction and the GC^{TS} structure as the transition state. Furthermore, the G^*C^* tautomer is a minimum in the approximation of a planar geometry of the bases, i.e. the C_S point group. In the vibrational spectra of the G^*C^* tautomer and the transition state structure there is also a vibrational mode with a harmonic frequency of about -30 cm^{-1} and components only normal to the plane of the base pair, which indicate both structures not to be planar in their true minima. The error, introduced by ignoring this fact, should be small, e.g. expressed in energy terms below 0.1 kcal mol^{-1}, which is far below the error introduced by the computational method.

Furthermore, a gap in the vibrational spectra of the GC base pair and the transition state structure is evident between 1900 cm^{-1} and 3100 cm^{-1}. In the vibrational spectrum of the G^*C^* tautomer two vibrations are shifted in this region to 2884 cm^{-1} and 2985 cm^{-1}, which have no direct equivalent in the spectra of the other two molecular complexes GC and GC^{TS}. Both vibrations are dominated by the protons Ha and Hb involved in the DPT reaction, which is further expressed by the reduced mass of 1.09 amu. Especially the vibration with the harmonic frequency of 2985 cm^{-1} is predicted to have a very strong infrared absorption, i.e. it is the strongest infrared absorption of the whole spectrum. Also in the spectrum of the transition state structure two vibrations exist, which seem to have no equivalent in the other two spectra. These vibrations have harmonic frequencies of 1776 cm^{-1} and 1893 cm^{-1} and moderate infrared activities. Both vibrations are the more interesting, since their reduced masses of about 1.70 amu imply that in these vibrations also different heavy atoms take part. The last fact is also consistent with the shortening of all H-bonds in the transition state structure discussed earlier in this section. Careful inspection of the vibrational spectrum of the GC base pair reveals two vibrations at 3112 cm^{-1} and 3210 cm^{-1} with moderate infrared activities. These vibrations are also dominated by the Ha and Hb protons and both vibrations have a reduced mass of about 1.09 amu. In the other two spectra no

Table 12.9 The Absolute energy E, the basis set superposition error BSSE, the zero-point vibrational energy ZVPE and the absolute entropy S^{298} at 298 K are given as calculated by the B3LYP/6-31G** approach for the isolated bases (G,C), their tautomers (G^*,C^*), the corresponding Watson-Crick base pairs (GC, G^*C^*) and the associated transition state of the DPT (GC^{TS})

molecule	absolute energy H	BSSE kcal mol^{-1}	ZPVE kcal mol^{-1}	S^{298} cal mol^{-1} K^{-1}
G	-542.5642016		72.8	88.0
G^*	-542.5635787		73.0	87.1
C	-394.9413506		61.5	79.1
C^*	-394.9393428		62.4	79.8
GC	-937.5549753	4.8	136.8	131.1
GC^{TS}	-937.5325383	5.0	132.3	119.2
G^*C^*	-937.5402094	4.6	136.5	122.6

equivalent vibration exists in this range of frequencies. All described vibrations have several properties in common. They are all infrared active, they have all the same symmetry, they are all dominated by the Ha and Hb protons and all have no direct equivalent in the other spectra. Most interesting is that the difference of the two vibrations in each spectrum is about 100 cm^{-1}. Therefore it is plausible to state that the vibrations are connected to each other, i.e. the two vibrations of the the GC vibrational spectrum are shifted to lower energies in the G^*C^* tautomer by 225 cm^{-1}. In the transition state structure there is an even larger shift, compared with the spectrum of the GC base pair, of more than 1300 cm^{-1}. These two vibrations also seem to play an important role in the DPT reaction.

The absolute values of the zero-point vibrational energy (ZPVE) of the isolated bases and the molecular complexes are summarized in tab. 12.9. For the reactions the relative ZPVEs of the products and educts are relevant, which are summarized in tab. 12.10. As in the AT base pair, the bases G and C have a smaller ZPVE than the GC base pair, and as a consequence ΔZPVE(GC) of the association reaction, G+C $\rightarrow$ GC, equals 2.5 kcal mol^{-1}. In case of the second association reaction, $G^*+C^* \rightarrow G^*C^*$, it is the other way round than in the AT base pair. Here the G^*C^* tautomer has a higher ZPVE than the sum of the ZPVEs of the individual bases. Therefore ΔZPVE(G^*C^*) equals 1.1 kcal mol^{-1}. Just like in the AT base pair the tautomeric structure (G^*C^*) has a smaller ZPVE than the base pair (GC) and ΔZPVE(taut) equals -0.3 kcal mol^{-1}. Furthermore, the transition state structure has the smallest ZPVE of all three molecular complexes listed in tab. 12.9, i.e. the ZPVE of the transition state structure is by 4.4 kcal mol^{-1} smaller than the ZPVE of the GC base pair, reducing the hight of the barrier in the

Table 12.10 Given are the relative energy ΔE^{B3LYP}, the difference ΔBSSE, the relative zero-point vibrational energy ΔZPVE, ΔE^0 the relative energy corrected for ΔZPVE and ΔBSSE and the relative entropies ΔS^{298} of the association reactions and the tautomerization. Given is for each quantity the difference of the sum of the products and the sum of the educts.

reaction	ΔE^{B3LYP} kcal mol^{-1}	ΔBSSE kcal mol^{-1}	ΔZPVE kcal mol^{-1}	ΔE^0 kcal mol^{-1}	ΔS^{298} cal mol^{-1} K^{-1}
G+C $\rightarrow$ GC	-31.0	4.8	2.5	-23.7	-36.0
G*+C* $\rightarrow$ G*C*	-23.4	4.6	1.1	-17.7	-44.3
GC $\rightarrow$ GC^{TS}	14.1	0.2	-4.4	9.9	-11.9
G*C* $\rightarrow$ GC^{TS}	4.8	0.4	-4.2	1.0	-3.5
GC $\rightarrow$ G*C*	9.3	-0.2	-0.3	8.8	-8.5

reaction GC $\rightarrow$ G*C* by a significant amount.

As in the case of the AT system, the absolute and relative entropies at 298 K ΔS^{298} are presented in tabs. 12.9 and 12.10. Both association reactions have large negative relative entropies, i.e. ΔS^{298}(GC)=-36 cal mol^{-1} K^{-1} and ΔS^{298}(G*C*)= -44 cal mol^{-1} K^{-1}. Both are comparable in magnitude with ΔS^{298}(AT) and ΔS^{298}(A*T*). Also the tautomerization reaction has a negative relative entropy (ΔS^{298}(taut)=-8.5 cal mol^{-1} K^{-1}), but the relative entropy is about three times as large as in the AT system. The transition state structure has the lowest absolute entropy S^{298} of the three molecular complexes, which is by almost 12 cal mol^{-1} K^{-1}), about one third to one fourth of ΔS^{298}(GC) or ΔS^{298}(G*C*), smaller than the absolute entropy of the GC base pair.

Energies and enthalpies The absolute energies as calculated by the B3LYP/6-31G** approach of the optimized structures of the GC base pair, the G*C* tautomer and the associated GC^{TS} transition state structure are summarized in tab. 12.9. The relative energies of the different reactions are given in tab. 12.10.

Here, the focus will be put first on the interaction energies of the molecular complexes and especially on the GC base pair, since an experimental estimate of the interaction enthalpy of the GC base pair exists [132], which predicts ΔH^{298}=-21.0 kcal mol^{-1}. The pure difference between the energies, calculated in the B3LYP/6-31G** approach, of the isolated bases (G and C) and the base pair (GC) amounts to -31.0 kcal mol^{-1}, which is by far too negative. The interaction energy can be further refined by accounting for the well-known BSSE, which has been determined in the standard counter-poise scheme of Boys and Benardi [138],

cf. tab 12.9. The BSSE of the GC base pair is determined to be 4.8 kcal mol^{-1}. After correcting the interaction energy for the BSSE it is estimated to be -26.2 kcal mol^{-1}. As a next step the difference of the ZPVEs has to be introduced (cf. eq. 12.1). Now there are now several estimates of ΔZPVE(GC) for the formation of the GC base pair. The calculation of Florian et al. [141] at the HF/MINI-1 level of theory results in a difference in the ZPVEs of 2 kcal mol^{-1}, whereas Gould and Kollman [124] refer to a value of 2.6 kcal mol^{-1} at 298 K in the HF/6-31G* approach. Furthermore, the empirical force field calculations of Del Bene suggest a difference in the ZPVEs of 2.4 kcal mol^{-1}. The B3LYP/6-31G** approach currently under investigation calculates the difference ΔZPVE(GC) as 2.5 kcal mol^{-1}, in good agreement with previous calculations. The corrected interaction energy of the GC is then given as -23.7 kcal mol^{-1}. The difference to the experimental estimate is still 2.7 kcal mol^{-1}, which is by far too large to be accounted for by a temperature correction. Actually to high an interaction energy of the GC base pair has already been expected in the previous subsection based on the underestimated bond length of the O10-N8 H-bond.

Similar problems have also been encountered by Gould and Kollman [124], who calculated the Hartree-Fock energy using the double-ζ basis set DZP on a HF/6-31G* optimized geometry of the GC base pair. This approach estimates the interaction energy to be -22.3 kcal mol^{-1} without any temperature correction. But, this approach is altogether questionable, since the intermolecular distances are underestimated as previously reported and the effects of electron correlation are completely neglected, although they are known to be very important for H-bonding [130]. Furthermore, the interaction energy is increased to 28.0 kcal mol^{-1} when the energies are determined at the MP2/DZP level of theory on the HF/6-31G* optimized structure. This is the opposite of what has been desired, but is in accordance with the too short H-bonding distances.

After the assessment of the accuracy that can be expected from the B3LYP/6-31G** approach for the GC system, the focus will be turned to the association reaction of the G*C* tautomer. The pure energy difference is -23 kcal mol^{-1}, which has to be corrected for the BSSE of 4.6 kcal mol^{-1} and the difference of the ZPVE of 1.1 kcal mol^{-1} to obtain the Interaction energy. The interaction energy of the G*C* tautomer at 0 K in gas phase is, therefore, the estimated to be ΔE^0= -18 kcal mol^{-1}. Florian et al. calculated in the HF/MINI-1 approach a smaller value of -18 kcal mol^{-1}, which is surely consistent with the present estimation. Furthermore, it should be noted that there is no extreme energy difference between the two association reactions of the GC system, as in the case of the AT system.

Now a closer look to the principle target of the present investigation, the tautomerization reaction, is taken. First of all, it was only possible to find one tautomeric structure. A minimum for the second, in principle possible, tautomer involving the H-bond O10-N8 and N11-O7 could not be located. This is consis-

tent with the observations of the calculations of section 11.2. The calculation of Florian et al. [131] did not find a second tautomer, too.

The uncorrected energy difference of the GC base pair and the G^*C^* tautomer is 9.3 kcal mol^{-1} at the B3LYP/6-31G** level of theory. After correction for the BSSE (-0.2 kcal mol^{-1}) and the difference of the ZPVEs (-0.3 kcal mol^{-1}) the interaction energy is reduced to ΔE^0=8.8 kcal mol^{-1}. This is in contrast to the findings of the HF/MINI-1 of Florian et al. [141], which predict an uncorrected energy difference of 0.5 kcal mol^{-1}. The correction for the BSSE and ΔZPVE reduces the value for ΔE^0 to -0.2 kcal mol^{-1}, i.e. the G^*C^* is predicted to be slightly more stable than the GC base pair, which is surely unphysical.

Comparing the results in the frozen core approximation (cf. section 11.2), which predict the energy difference to be 34 kcal mol^{-1}, to the present estimation of ΔE^0 (=8.8 kcal mol^{-1}), the energy difference is largely reduced. It is most important that the energy difference ΔE^0 of the GC base pair and the G^*C^* tautomer is smaller than the energy released by the pairing of the G and C bases, which may be the source of the energy necessary to drive the DPT reaction.

Concerning the corrected energy of the GC^{TS} structure the present B3LYP/6-31G** approach predicts an energy difference ΔE^0 to the GC base pair of 9.9 kcal mol^{-1}. In case of the G^*C^* tautomer the energy difference to the transition state structure is predicted to be only 1.0 kcal mol^{-1}, tripling the previous estimates of Florian et al. [141]. The importance of this finding cannot be overestimated, since it suggests that the tautomeric structure lives significantly longer than one vibrational period. Furthermore, just as in the case of the AT base pair the relative energy of the transition state structure is lowered as the two bases appoach each other. So, when the bases are bound in a double helix, which is quite stiff, it is probably not possible for the bases to come that close to each other. This in turn means that the relative energy of the transition state structure in a double helix is larger and the average lifetime of the tautomeric structure is significantly longer.

12.4 Discussion

In a discussion of molecular properties it is always desirable to connect the computed data to experimentally accessible quantities, in order to assess the obtained accuracy of the results. Two such connections have already been presented, the X-ray structure of the AU base pair [133] and GC base pair [140] and the strength of the H-bonds [132]. Another quantity, for which abundant experimental data are available in the literature, is the rate of spontaneous point mutations of different DNAs, which have been determind by experiments [142, 143] to be in the range of 10^{-6} to 10^{-10}.

Löwdin connected the rate of spontanous point mutations in the 1960's to the occurence of a certain amount of base pairs in their stable tautomeric form,

e.g. A^*T^* and G^*C^* [40, 95]. The tautomeric base pairs may lead to errors in the replication process and subsequently a point mutation is introduced in the DNA copy. Therefore, it is of value to know the equlibrium constant for the tautomerization reaction in the AT and GC base pair. In standard thermodynamics the equilibrium constant K is directly connected to the Gibb's free enthalpy by:

$$K = exp\left\{-\frac{\Delta G}{RT}\right\} \tag{12.3}$$

with R the gas constant and T the temperature. Additionally, an estimate of the rate constant k is possible for the reactions whose transition state is known. Since the entropy and the heat capacity are not constant in the temperature range from 0 K to 298 K it is not sufficent to evaluate Gibb's free enthalpy as $\Delta G^T = \Delta E^0 + T\Delta C_v - T\Delta S$, but it has to be evaluate:

$$\Delta G^T = \Delta E^0 + \int_0^T \Delta C_v(T')dT' - \int_0^T \Delta S(T')dT' \tag{12.4}$$

with $\Delta C_v(T')$ the difference of heat capacities at constant volume and $\Delta S(T')$ the difference of the entropies. This is still an approximation, since any expansion work is neglected. The error introduced is below 1 kcal mol^{-1}. As a byproduct the enthalpy ΔH^{298} is obtained simultanously for a direct comparison with the gas phase association experiments [132]. Concerning the association reaction of the AT base pair, the observed ΔH^{298} of ref. [132] is some mixture of the interaction enthalpy of the Watson-Crick type and Hoogsteen type AT base pair. This is evident from the calculations of Gould and Kollman [124], which refer to a difference in the interaction enthalpy ΔH^{298} of the Watson-Crick type and Hoogsteen type AT base pair of 1 kcal mol^{-1} or less over a range of quantum chemical models, which always predict the Hoogsteen type base pair to be the more stable one. Since 1 kcal mol^{-1} is about one to two times RT in the experimental temperature range investigated in ref. [132] both species are formed. In tab. 12.11 the enthalpies ΔH^{298}, the free enthalpies ΔG^{298} and the corresponding equilibrium constants (or rate constants) at 298 K are summerized.

Concentrating on the DPT reactions in the AT and GC base pair, it is evident that both equilibrium constants are well inside the range of the experimental point mutation frequency, but tend to the lower end of the range. Furthermore it seems to be more likely to achieve the DPT reaction in the GC base pair than in the AT base pair by a factor of 10. Here it may be appropriate to emphasize the error in the prediction of the equilibrium constants. In the calibration of the applied B3LYP/6-31G** method in the preceeding subsections an error in the predicted energy of about 2 kcal mol^{-1} has been revealed, i.e. about 3 to 4 times RT at T=298 K. This corresponds to an error of a factor of 20 to 50 in the equilibium

Table 12.11 The enthalpy ΔH^{298}, free Gibb's enthalpy ΔG^{298} and the equilibrium constant K^{298} are given at 298 K as calculated by the B3LYP/6-31G** approach. The equilibrium constant of reactions with a transition state as products denote the rate constant of the corresponding reaction. A direct comparison of the predicted ΔH^{298} values with those of the experiment [132] is allowed. The equilibrium constants published so far [132], are not comparable to the calculated data, since they still contain apparative constants.

reaction	ΔH^{298} kcal mol^{-1}	ΔG^{298} kcal mol^{-1}	K^{298}
A+T $\rightarrow$ AT	-10.4	0.2	0.72
A*+T* $\rightarrow$ A*T*	-24.9	-13.6	3×10^{12}
AT $\rightarrow$ A*T*	11.2	11.9	5×10^{-9}
G+C $\rightarrow$ GC	-23.3	-17.4	7×10^{13}
G*+C* $\rightarrow$ G*C*	-17.8	-4.6	2×10^{3}
GC $\rightarrow$ G*C*	8.1	10.6	4×10^{-8}
GC $\rightarrow$ GC^{TS}	8.8	12.3	2×10^{-9}
G*C* $\rightarrow$ GC^{TS}	0.6	1.7	6×10^{-2}

constants, which furthermore implies that at least the order of magnitude should be correct. The estimates made are valid for the gas phase without any additional sources of energy. In a real DNA system the association reaction of the considered base pair or its neighbours may offer additional energy to the system, which may drive the DPT reaction. Especially the energy release of the GC pairing reaction is large enough to facilitate the DPT reaction. This will largely enhance the amount of tautomers. In essence it may be stated that based on the predicted equilibrium constant it is about equally propable to find an A*T* and G*C* tautomer and it is a possible pathway to a point mutation, although not the only one.

Further support to the mutation hypothesis is given by the small energy difference of the association reaction of the GC base pair, G+C $\rightarrow$ GC (ΔH^{298}=-23 kcal mol^{-1}), and the association reaction of the A*T* tautomer, A*+T* $\rightarrow$ A*T* (ΔH^{298}=-25 kcal mol^{-1}), i.e. the H-bonds of the base pairs are of about the same strength. It is easy to imagine that an A*T* tautomer is misinterpreted as a GC base pair, because the opening of the DNA double strand and subsequently the recognition of the base, as well as some error recognition mechanisms are believed to depend on the interaction energy of the base pairs [144–146].

In principle, ionic structures of the base pair like A$^+$T$^-$ and A$^-$T$^+$ (and the equivalents of the GC base pair) may be formed by a single proton tranfer in one of the H-bonds. The HF/MINI-1 investigation of Florian et al. [131] revealed

no local minimum of the PES for the ionic structures, but the investigation of the small 3-21G atomic basis set in connection with the B3LYP method even did not reveal a minimum for the A^*T^* tautomer, which definitely exist. This encouraged a new search for the ionic structure, but the attempt remaines unsuccessful. Neither for the AT mor the GC base pair any ionic structure could be located.

Furthermore, a tautomeric structure involving both asymmetric H-bonds of the GC base pair could not be found, which seems to be constistent with the NMR and calorimetry study [147] of the tautomerism in the GC base pair, formed by ribose modified nucleosides in non-polar solvents. This study revealed different rotation rates of the amino groups in the GC base pair about their N-C single bond. More precisely, the rotation of the amino group of the guanine occurred within the closed base pair, whereas the rotation of the amino group of the cytosine occurred only during the transient opening of the base pair. The authors considered both the double proton transfer forming the G^*C^* tautomer and the mutual polarization of the bases as a possible for the observed differences in the rotation dynamics. The proposed effect of mutual polarization of bases involved the formation of partial positive and negative charges on the N8 and N3 nitrogen atoms of cytosine, respectively. This distributes the electron density of the π-bond and the H-bonds more evenly over the bonds involved in the tautomerism, as a consequence of the H-bonding between the bases. This effect was supposed to increase the bond order of the amino bond of cytosine, which would to slow down the rotation rate of this group. In the subsequent variable temperature IR study of the same group [148] the conclusion was drawn, that the probability of the G^*C^* tautomerism is not sufficient to explain the observed differences in the rotation dynamics. However the authors did not provide any estimate of the sensitivity of their technique to the detection of an amount smaller than 1% of the G^*C^* tautomer in the spectra. It should be moted that a G^*C^* concentration in the experimentally inaccessible range below 10^{-3} may, in principle, cause a biologically significant frequency of spontaneous point mutations. The results of the present B3LYP/6-31G** approach do not confirm the predictions of the 'mutual polarization' mechanism.

From the viewpoint of the current results, the different rotation dynamics of the amino groups in the GC base pair, may be understood in a different way. First it may noted that even in the free guanine the C2-N11 bond is longer than the C4-N8 bond in the free cytosine, indicating a larger single bond character of the C2-N11 bond of the guanine. This difference is not influenced by the H-bonding in the GC base pair and amounts to 0.014 Å. Second, the N8-O10 H-bond is significantly shorter than the O7-N11 H-bond. The shorter H-bond indicates a higher H-bonding energy for the N8-O10 bond, which makes the amino group of the cytosine more rigid than the guanine one. Finally, the transfer of the Ha proton is energetically more feasible than the transfer of the Hc proton. In view of

the long time scale of the NMR measurements (in comparison with the time scale of molecular vibrations) only averaged quantities are observed, which leads to a higher observed bond order of the C4-N8 bond.

The mutation hypothesis suggested by Löwdin has been severely critized in the past [96], since in order to become operative a certain amount of the base pairs have to be assumed in their tautomeric form during the replication. Löwdin suggested [94, 107] an explanation with a proton tunnelling process through the energy barrier between the AT base pair and the A^*T^* tautomeric structure. The presented quantum chemical model reveals for the first time a significant energy difference between the base pair (AT, GC) and the imino-enol tautomer (A^*T^*, G^*C^*) and a fairly high lying transition structure of about 3 kcal/mol (cf. tab. 12.5). This accounts for a lifetime of the imino-enol tautomer being longer than one stretch-vibration.

13 Outlook

Where may one go from here? Certainly, the presented calculations are not at the end of their ability nor at the end of the present computer resources. One major lesson to learn from part I is that it is crucial to take all biological information into account that are available. Otherwise, artificial information may rapidly be gained which is prone to misinterpretation. An extreme example is the interpretation of DNA base sequences in the frame of fractals (see section 3). Here it was tried to deduce form the 'degree' of fractality, i.e. the parameter α, whether a sequence is protein coding or non-coding [39]. It has been shown, that the fractality of a given base sequence is directly linked or caused by the base pair composition of the sequence and the fluctuations in the base pair composition [56, 57]. So the fractality or better the 'pseudo fractality' is not directly linked to the coding or non-coding character of a given base sequence. Nevertheless, there are of course statistical differences between protein coding and non-coding base sequences as has been known for some time now. These differences can be detected with the method presented in section 3.1.4 by the comparison of the 'pseudo-fractal' parameters of a natural base sequence and the 'pseudo-fractal' parameter of associated artificial computer generated base sequences [59, 81]. Altogether it does not seem to be very fruitful to continue with highly averaging methods, but it is necessary to introduce biological information gained elsewhere. One example is given in section 8, where only protein coding sequences are considered and furthermore the triplett structure of the codons is taken into account. Introducing further biological information to the sequence selection for the averages leeds to a more pronounced result. Certainly, this shows the path to follow and it may be questioned whether it is important for the translation process that the famous 5'-TATA-3' box, which indicates the start of a reading frame, shows a QNQ pattern.

Finding a path to follow with the statistical methods, there is still question why the groups of consecutive bases Q, N and R are chosen as they were in section 8. Or why should consecutive bases be called quantum entangled or quantum correlated when two protons come close to each other, i.e. 3 Å to 4 Å. First of all the findings presented in section 8 show that the choice of the groups Q, N and R is legitimate. For the question why some consecutive bases are called quantum entangled (for short time scales) classical quantum chemical program packages may give advice finding the PES in which the protons are moving and showing the area of real space accessible to the protons. Of course these calculations will involve the determination of rather large molecular system of 60 and more atoms with the inclusion of an estimation of the effects of electron correlation. This immediately prevents simple Hartree-Fock calculations from being useful. With the current quantum chemical program packages these calculations should be possible without too much effort, since no determination of a transition state is necessary,

which is still difficult to determine in a large molecular system under inclusion of the electron correlation effects. This had already affected the calculations of the relatively 'simple' AT base pair of the Watson-Crick type, which consists of 30 atoms. Here, difficulties have been encountered during the search for the transition state of the double proton transfer. The molecular system was drifting in real space instead of converging to a stationary point. So certainly the geometry optimizer of the quantum chemical program packages have to be improved.

Also in the case of the 'simple' Watson-Crick type base pairs to step on, e.g. the reaction path of the double proton transfer could be followed. In the following it may be judged whether the assumption of the transfer of the protons along the connecting line of the heavy atoms participating in the H-bonds during the double proton transfer is reasonable (cf. secton 11.1).

There are several other points to investigate, e.g. whether metastable or resonant states do exist in the double proton transfer. But most important, after all theoretical work, are experiments. I will suggest two, which should be relatively easy to perform with some care.

13.1 Proposed experiments

First I propose to measure the reaction rate constants of DNA double strand association k_a and dissociation k_d for a certain DNA oligomer as a function of the deuterium content x_D (x_D=[D]/([H]+[D])) of the buffer solution in use. Other important parameters should be approximately physiological, i.e. the temperature could be T=25°C, the pH of the solution should be approximately 7 and the pressure should be 1 bar. All these parameters have to be constant for the experiments with different x_D to be meaningful.

The oligomer to be investigated should fulfill some requirements on the expected k_a and k_d, in a purely hydrogenous medium, i.e. x_D=0. In order to be easily measureable the rate constants should not be larger than 10^7 $M^{-1}sec^{-1}$ and not smaller than 1 $M^{-1}sec^{-1}$. This can be achieved with different DNA oligomers of 10 to 14 monomers which often have a k_a in the range of 10^5 to 10^6 $M^{-1}sec^{-1}$ with increasing number of H-bonds between the two strands, and a k_d of 1 to 10^3 $M^{-1}sec^{-1}$, with a decreasing number of H-bonds between the two strands [2]. Furthermore, the oligomer has not to be a palindrome to prevent self association of the single strand DNA. Another, but not that important, point is that the oligomer should not incorporate matching subsequences, i.e. if the subsequence 5'-AGAT-3' does exist the subseqence 5'-ATCT-3' should not appear in the oligomer.

Next the connection to section 8 is presented. The idea is to find an oligomer with as many as possible so called quantum entangled consecutive bases as well as an oligomer with as few as possible such bases. Let me point out here that

when the oligomer 5'-xyz-3' maximizes the number of quantum entangled consecutive bases the reverse oligomer 5'-zyx-3' immediatly minimizes this number. Therefore I propose to investigate the oligomer 5'-GCCTGCTCTAA-3' with a high number of quantum entangled consecutive bases and 5'-AATCTCGTCCG-3' with a small number.

Before the expected results are described some of the experimental details are given. The experiment starts with two single stranded DNA oligomers. One oligomer is chemically bonded to an optic fiber which is a standard procedure. A dye, e.g. fluorescein, is bonded to the second oligomer. These modified oligomers are commercially available. The DNA modified optic fibre is placed in a tube which is illuminated by a light appropriate to promote the dye to its electronic excited state. Next the optic fibre and the second DNA oligomer are incubated for an hour in a 20mM phosphate buffer with the deuterium content x_D^{test} to be investigated. This time is sufficient to equilibrate all exchangeable proton positions with the deuterium of the buffer [2], but the fluorescence spectrum of the fluorescein will remain unchanged in shape and position. Next the buffer with the second DNA oligomer is put at constant flow to the reaction tube and the fluorescence coupled to the fibre is recorded with time. From these data the rate constant of association k_a is extracted. When the maximal fluorescens intensity is reached the buffer is switched to a buffer without DNA and $x_D = x_D^{test}$ and the fluorescens intensity is recorded, from which the dissociation rate constant k_d is extracted. Here only the principle of the apparatus is described; the real implementation is rather complex and is described in detail in ref. [149], where also the detail of the data analysis is given.

Additionally a short recipe is given for a buffer with a well-known deuterium content x_D and a defined pH*[7], since it is not a standard procedure. One of the major experimental problems here is that it is impossible to measure pH* in a deuterium containing solution with a normal pH-meter. The strategy to overcome this problem is to dilute a purely hydrogenous buffer of a high concentration, e.g. a 400mM phosphate buffer, which is calibrated to the correct pH to the appreciated buffer concentration in the experiment, here a 20mM poshate buffer, with a H_2O-D_2O mixture of appropriate x_D. Therefore it is necessary to control the masses of H_2O and D_2O used precisely with a balance, at least up to one mg. One drawback to this procedure is that no purely deuterated buffer is possible. In the example put forward the limit is a buffer with x_D=0.95, which may even be driven a little further, but x_D=1 will never be reached.

In the end I will give some thoughts on the expected results. It is expected that quantum entanglement may preferably exist beween identical particles. So when

[7] In this context I will allways adress with pH* the negative logarithm of the combined concentration of H^+ and D^+ ions, i.e. pH*=-$\log_{10}$([H^+]+[D^+]) and for a solution with x_D=0 pH*=pH.

one of the protons in a quantum entangled pair is exchanged with a deuteron the pair may not any longer be quantum entangled and looses its properties due to quantum entanglement. Also exchanging the second proton in the pair quantum entanglement may be possible again, but the mass of the pair will be approximately doubled, which influences the rate constants. Here it should be kept in mind that for some quantum entangled proton pairs possibly only one proton may be exchanged, i.e. also in a purely deuterated medium the quantum entanglement of this pair may not be recovered.

So what could be the influence of quantum entanglement on the rate constants? In the analysis the all-or-none mechanism will be assumed for simplicity, as is often done when studying kinetics of DNA [2]. This allows to postulate for the highly complex process an Arrhenius type formula for the rate constants k_a and k_d

$$k = A^{eff} \times exp\{-E_a^{eff}/RT\} \tag{13.1}$$

with R the gas constant and T the temperature. The superscript eff indicates that the observed quantities are always effective ones, but the usual meaning is preserved. E_a^{eff} represents the activation energy and A^{eff} the collision coefficient, which is a measure for the number of reactive encounters per time unit. It is not expected that the activation energy E_a^{eff} is altered by the presence or absence of quantum entanglement. Nevertheless care should be taken of the activation energy for the protonated species $E_a^{eff,H}$ and the deuterated species $E_a^{eff,D}$ which may differ. In a H,D-mixture E_a^{eff} is expected to be an intermediate depending on the D replacement in the relevant H-bonds. In order to test these assumptions it would be important to measure the temperature dependence of the rate constants k_a and k_d, too. But what is expected to be affected by quantum entanglement is the collision coefficient A^{eff}, since the symmetry of the density matrix of the protons is altered. This is important in the collision process as well as in the question whether the collision is a reactive one or not. So in essence I do expect k_a to be similar in the purely hydrogenated and purely deuterated case. The same is true for k_d. But in case of a mixture, say x_D=0.5, I expect k_a and k_d to be significantly different from the former ones. Precisely, I expect the k_a (or k_d) for the mixture to be larger or smaller than the k_a's (or k_d's) of the pure solutions. Up to now it is not clear whether the symmetry brought to the system by quantum entanglement supports the reaction, i.e. enlarges A^{eff}, or hinders the reaction, or most interesting, depends on the sequence investigated. Another interesting point about this behaviour of k_a (or k_d) vs. x_D is that an interpretation in the frame of fractionation theory is immediately prevented, since the fractionation theory assumes that the rate constant in the mixture has to remain in the range of the rate constants for purely hydrogenated and purely deuterated solutions. One last point is that the pH^* for mixtures is reproducible but not very well known. So the

dependence of k_a and k_d on pH* should be investigated which I expect to be flat around pH*=7.

Summarizing I propose to measure the rate constants of double strand association k_a and dissoziation k_d of different DNA oligomers, e.g. 5'-GCCTGCTCTAA-3' and 5'-AATCTCGTCCG-3', in dependence of the deuterium content of the buffer solution x_D, the combined H^+ and D^+ concentration called pH* and the temperature T.

The second experiment is similar to the DNA experiment, but focuses on another class of biomolecules, the enzymes. For the experiments I have selected the glucose oxidase (GOD), which is often isolated from *Aspargillus niger*. GOD catalyzes the oxidation of glucose to δ-gluconolactone and the subsequent reduction of oxygen to hydrogen peroxide. The enzyme contains one flavin adenine nucleotid (FAD) cofactor per monomer and is a homodimer with a molecular wight of 132-320 kDa, depending on the extent of surface glycosylation. The GOD is a well studied, robust and commercially available enzyme. Furthermore the mechanism of the GOD is known to make use of proton tunnelling [16], so the quantum nature of the proton is important for the process. The largest kinetic isotope effect on the Arrhenius prefactors has been observed for the lightest GOD species [16], i.e. with the smallest amount of surface glycosylation. These experiments have been carried out in a purely hydrogenous medium with [1-D]-2-desoxyglucose in which the deuterium is covalently bonded to a carbon atom. In contrast to this, in the experiments which are be proposed only the exchangeable protons of the glucose and the enzyme may by deuterated.

Here I propose to measure the specific activity or turnover number of GOD for beta-D-glucose in the presence of benzoquinone as a function of the deuterium content x_D of the buffer medium. The pH* dependence should be investigated around the pH* optimum of GOD in the hydrogenous medium of 5.5. Later on also other kinetic properties of GOD, like the Michaelis-Menten constant K_M or k_{cat} as well as their temperature dependence are of interest. For the result I do expect the turnover number in the buffer with $x_D \approx 1$ to be lower by a factor of 1.5 to 2 than in the buffer with x_D=0. But the turnover number in the buffer with $x_D \approx 0.5$ is expected to be even smaller. This is also indicated by first preliminary experiments.

14 Acknowledgements

I wish to express my deepest gratitude to Professor Dr. Chatzidimitriou-Dreismann for his encouriging support of my work an for many fruitful discussions.

I would like to acknowledge the support of the TU Berlin in the frame of the ”Beschäftigungsplanungmittel” for 6 months.

I wish to thank Prof. Dr. Renger for his support.

Furthermore, I wish to thank Tyno, for many coffee-breaks, discussions, and many hours he spend with me in one room.

For his support, encouragement, his introduction a more biological view of things and some wonderfuls visits to Uppsala I wish to thank Professor Dr. Larhammar. Unforunately I havn’t seen you for a while now!

I would like to acknowledgement the support I gained through the complete INSI team, because I am not able to list them all. I will pick only two.
Professor Dr. Findenegg I wish to thank for one year of employment and Dr. Steitz for his patience.

Professor Dr. Karlsson I wish to thank for several experiment at the eVS instrument, discussions, and his reminder of how science works.

Hey, Samy and Manuela, thank you for many beers, games and evenings I haven’t thought of science.

Irmelin and Michel, would any one be able to read this without your corrections?

Bärbel and Heinz, what would I have done without your support!

Christine and Christoph I wish to thank for many hours I was late and all the support I gain from them.

In the end I wish to thank Katrin, for all my scientific journeys, her never ending encouragement and all the things I do forget now!

References

[1] J. D. Watson and F. H. C. Crick, Nature **171**, 964 (1953).

[2] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry Part III: The behavior of biological macromolecules* (W. H. Freeman, San Francisco, 1980).

[3] M. F. Perutz, Nature **228**, 726 (1970).

[4] G. R. Jacobson and G. R. Stark, *The Enzymes vol. 9* (Academic Press, New York, 1973).

[5] O. Carnal and J. Mlynek, Phys. Rev. Lett. **66**, 2689 (1991).

[6] D. W. Keith, C. R. Ekstrom, Q. A. Turchette, and D. E. Pritchard, Phys. Rev. Lett. **66**, 2693 (1991).

[7] M. S. Chapman, C. R. Ekstrom, T. D. Hammond, R. A. Rubenstein, J. Schmiedmayer, S. Wehninger and D. E. Pritchard, Phys. Rev. Lett. **74**, 4783 (1995).

[8] M. Arndt, O. Nairz, J. Voss-Andreae, C. Keller, G. Van der Zouw and A. Zeilinger, Nature **401**, 680 (1999).

[9] E. F. Caldin, Chem. Rev. **69**, 135 (1969).

[10] A. J. Kresge and M. F. Powel, J. Am. Chem. Soc. **103**, 201 (1981).

[11] L. Reocker and T. J. Meyer, J. Am. Chem. Soc. **109**, 746 (1987).

[12] Y. Cha, C. J. Murray, and J. P. Klinman, Science **243**, 1325 (1989).

[13] M. M. Palcic and J. P. Klinman, Biochemistry **22**, 5957 (1983).

[14] K. L. Grant and J. P. Klinman, Biochemistry **28**, 6597 (1989).

[15] M. E. Schneider and M. J. Stern, J. Am. Chem. Soc. **94**, 1517 (1972).

[16] A. Kohen, T. Jonsson, and J. P. Klinman, Biochemistry **36**, 2603 (1997).

[17] T. Jonsson, M. H. Glickman, S. Sun, and J. P. Klinman, J. Am. Chem. Soc. **118**, 10319 (1996).

[18] M. H. Glickman and J. P. Klinman, Biochemistry **34**, 14077 (1995).

[19] H. Sumi and J. Ulstrup, Biochim. Biophys. Acta **955**, 26 (1988).

[20] J. K. Hwang, Z. T. Chu, A. Yadav, and A. Warshel, J. Phys. Chem. **95**, 8445 (1991).

[21] J. K. Hwang and A. Warshel, J. Am. Chem. Soc. **118**, 11745 (1996).

[22] L. Melander and W. H. Saunders, *Reaction Rates of Isotopic Molecules* (Wiley, New York, 1980).

[23] E. Guilini, E. Joos, C. Kiefer, J. Kupsch, I.-O. Stamatescu and H. D. Zeh, *Decoherence and the Apperence of a Classical World in Quantum Theory* (Springer Verlag, Berlin, 1996).

[24] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wiemann and E. A. Cornell, Science **269**, 198 (1995).

[25] M. R. Andrews, C. G. Townsend, H.-J. Miesner, D. S. Durfee, D. M. Kurn and W. Ketterle, Science **275**, 637 (1997).

[26] C. A. Chatzidimitriou-Dreismann, U. K. Krieger, A. Möller, and M. Stern, Phys. Rev. Lett. **75**, 3008 (1995).

[27] C. A. Chatzidimitriou-Dreismann, T. Abdul-Redah, R. M. F. Streffer, and J. Mayers, Phys. Rev. Lett. **79**, 2839 (1997).

[28] C. A. Chatzidimitriou-Dreismann, T. Abdul-Redah, and R. M. F. Streffer, Ber. Bunsenges. Phys. Chem. **102**, 544 (1998).

[29] E. B. Karlsson, C. A. Chatzidimitriou-Dreismann, T. Abdul-Redah, R. M. F. Streffer, B. Hjörvarsson, J. Öhrmalm and J. Mayers, Europhys. Lett. **46**, 617 (1999).

[30] G. I. Watson, J. Phys. Condens. Matter **8**, 5955 (1996).

[31] J. Sperling, H. Tributsch, R. M. F. Streffer, T. Abdul-Redah and C. A. Chatzidimitriou-Dreismann, J. Electroana. Chem. **477**, 62 (1999).

[32] E. Caldin and V.Gold, *Proton Transfer Reactions* (Chapman and Hall, London, 1975).

[33] V. Gold and M. A. Kessick, Dics. Faraday Soc. **39**, 84 (1965).

[34] V. Gold, Faraday Soc. **56**, 255 (1960).

[35] G. J. Kearley, F. Fillaux, M.-H. Baron, S. Bennington and J. Tompkinson, Science **264**, 1285 (1994).

[36] F. Fillaux, Physica D **113**, 172 (1998).

[37] I. Schlönvogt, S. Pitsch, C. Lesueur, A. Eschenmoser, B. Jaun and R. M. Wolf, Helv. Chim. Acta **79**, 2316 (1996).

[38] A. Eschenmoser and M. Dober, Helv. Chim. Acta **75**, 218 (1992).

[39] C. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. Stanley, Nature **356**, 168 (1992).

[40] P.-O. Löwdin, Adv. Quantum. Chem. **2**, 212 (1965).

[41] C. Gabriel, E. H. Grant, R. Tataa, P.R. Brown, B. Gestblom and E. Noreland, Nature **328**, 145 (1987).

[42] M. Frank-Kamenetskii, Nature **328**, 108 (1987).

[43] W. Li, Int. J. Bifurc. Chaos **2**, 137 (1992).

[44] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[45] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[46] P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature **360**, 636 (1992).

[47] J. Maddox, Nature **358**, 103 (1992).

[48] I. Amato, Science **257**, 747 (1992).

[49] P. Yam, Sci. Am. **267**, 13 (1992).

[50] W. Li and K. Kaneko, Nature **360**, 635 (1992).

[51] V. V. Prabu and J. Claverie, Nature **359**, 782 (1992).

[52] C. Chatzidimitriou-Dreismann and D. Larhammar, Nature **361**, 212 (1993).

[53] S. Karlin and V. Brendel, Science **259**, 677 (1993).

[54] S. Nee, Nature **357**, 450 (1992).

[55] B. Borstnik, D. Pumpernik, and D. Lukman, Europhys. Lett. **23**, 389 (1993).

[56] C. A. Chatzidimitriou-Dreismann, R. M. F. Streffer, and D. Larhammar, Biochim. Biophys. Acta **1217**, 181 (1994).

[57] C. A. Chatzidimitriou-Dreismann, R. M. F. Streffer, and D. Larhammar, Eur. J. Biochem. **224**, 365 (1994).

[58] D. Larhammar and C. Chatzidimitriou-Dreismann, Nucleic Acids Res. **21**, 5167 (1993).

[59] C. A. Chatzidimitriou-Dreismann, R. M. F. Streffer, and D. Larhammar, in *New Frontiers in Theoretical Biology*, edited by C. A. Dreismann (Hardronic Press, Palm Harbour, 1996).

[60] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).

[61] K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, Ma., 1949).

[62] E. Shannon, Bell. Syst. Tech. J. **27**, 379 and 623 (1948).

[63] E. Shannon and W. Weaver, *The mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, 1949).

[64] F. Flam, Science **266**, 1320 (1994).

[65] K. Konopka and C. Martindale, Science **268**, 789 (1995).

[66] C. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[67] S. M. Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons and H. E. Stanley, Biophys. J. **67**, 64 (1994).

[68] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1951).

[69] C. A. Chatzidimitriou-Dreismann, Int. J. Quantum Chem. **23**, 1505 (1983).

[70] W. Hendrix, J. W. Roberts, F. W. Stahl, and R. A. Weisberg, *Lambda II* (Cold Spring Harbor Lab. Press, Cold Spring NY, 1983).

[71] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier, Science **228**, 953 (1985).

[72] A. J. Jeffreys, V. Wilson, and S. L. Thein, Nature **316**, 76 (1985).

[73] G. Levinson and G. A. Gutman, Mol. Biol. Evol. **4**, 203 (1987).

[74] L. E. Orgel and F. H. C. Crick, Nature **284**, 604 (1980).

[75] C. P. L. Emerson and S. I. Bernstein, Annu. Rev. Biochem. **56**, 695 (1987).

[76] C. A. Chatzidimitriou-Dreismann and D. Seifert, Helv. Chim. Acta **81**, 584 (1998).

[77] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. A. Fields, J. D. Gocayne, J. D. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith and J. C. Venter, Science **269**, 496 (1995).

[78] S. G. Oliver, Q. J. M. Van der Aart, M. L. Agostini-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J. P. G. Ballesta, P. Benit, G. Berben, E. Bergantino, N. Biteau, P. A. Bolle, M. Bolotin-Fukuhara, A. Brown, A. J. P. Brown, J. M. Buhler, C. Carcano, G. Carignani, H. Cederberg, R. Chanet, R. Contreras, M. Crouzet, B. Daignan-Fornier, E. Defoor, M. Delgado, J. Demolder, C. Doira, E. Dubois, B. Dujon, A. Dusterhoft, D. Erdmann, M. Esteban, F. Fabre, C. Fairhead, G. Faye, H. Feldmann, W. Fiers, M. C. Francinques-Gaillard, L. Franco, L. Frontali, H. Fukuhara, L. J. Fuller, P. Galland, M. E. Gent, D. Gigot, V. Gilliquet, N. Glansdorff, A. Goffeau, M. Grenson, P. Grisanti, L.A. Grivell, M. de Haan, M. Haasemann, D. Hatat, J. Hoenicka, J. Hegemann, C. J. Herbert, F. Hilger, S. Hohmann, C. P. Hollenberg, K. Huse, F. Iborra, K.J. Indige, K. Isono, C. Jacq, M. Jacquet, C. M. James, J. C. Jauniaux, Y. Jia, A. Jimenez, A. Kelly, U. Kleinhans, P. Kreisi, G. Lanfranchi, C. Lewis, C. G. van der Linden, G. Lucchini, K. Lutzenkirchen, M. J. Maat, L. Mallet, G. Mannhaupt, E. Martegani, A. Mathieu, C. T. C. Maurer, D. McConnell, R. A. McKee, F. Messenguy, H. W. Mewes, F. Molemans, M. A. Montague, M. Muzi Falconi, L. Navas, C. S. Newion, D. Noone, C. Pallier, L. Panzeri, B. M. Pearson, J. Perea, P. Philippsen, A. Pierard, R. J. Planta, P. Plevani, B. Poetsch, F. Pohl, B. Purnelle, M. Ramezani Rad, S.W. Rasmussen, A. Raynal, M. Remacha, P. Richterich, A. B. Roberts, F. Rodriguez, E. Sanz, I. Schaaff-Gerstenschlager, B. Scherens, B. Schweitzer, Y. Shu, J. Skala, P.P. Slonimski, F. Sor, C. Soustelle, R. Spiegelberg, L. I. Stateva, H. Y. Steensma, S. Steiner, A. Thierry, G. Thireos, M. Tzermia, L. A. Urrestarazu, G. Valle, L. Vetter, J. C. van Vliet-Reedijk, M. Voet, G. Volckaert, P. Vreken, H. Wang, J. R. Warmington, D. von Wettstein, B. L. Wicksteed, C. Wilson,

H. Wurst, G. Xu, A. Yoshikawa, F. K. Zimmermann and J. G. Sgouros, Nature **357**, 38 (1992).

[79] J. Sulston, Z. Du, K. Tomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, S. Dear, A. Coulson, M. Craxton, R. Durbin, M. Berks, M. Metzstein, T. Hawkins, R. Ainscough and R. Waterston, Nature **356**, 37 (1992).

[80] B. Levin, *Genes V* (Qxford University Press, Oxford, 1994).

[81] C. A. Chatzidimitriou-Dreismann, R. M. F. Streffer, and D. Larhammar, Nucleic Acids Res. **24**, 1676 (1996).

[82] R. Omnès, *The Interpretation of Quantum Mechanics* (Princeton University Press, Princeton, 1994).

[83] R. Rein, P. Claverie, and M. Pollack, Int. J. Quantum Chem. **2**, 129 (1968).

[84] V. I. Poltev and B. I. Sukhorukov, Stud. Biophys. **24/25**, 179 (1970).

[85] R. Rein, Adv. Quantum Chem. **7**, 335 (1973).

[86] H. Fujita, A. Imamura, and C. Nagata, J. Theor. Biol **45**, 411 (1974).

[87] Z. G. Kudritskaya and V. I. Danilov, J. Theor. Biol **59**, 303 (1976).

[88] R. L. Ornstein, R. Rein, D. L. Breen, and R. D. Macelroy, Bioploymers **17**, 2341 (1978).

[89] M. Aida, J. Comput. Chem **9**, 362 (1988).

[90] C. Alhambra, F. Luque, F. Gago, and M. Orozco, J. Phys. Chem. **101**, 3846 (1997).

[91] Y. Nakamura, K. Wada, Y. Wada, H. Doi, S. Kanaya, T. Gojobori and T. Ikemura, Nucleic Acids Res. **24**, 214 (1996).

[92] R. M. F. Streffer, B. Schluricke, and C. A. Chatzidimitriou-Dreismann, paper in preparation (2001).

[93] H. Primas and U. Müller-Herold, *Elementare Quantenchemie* (Teubner, Stuttgart, 1984).

[94] Y. S. Kong, M. S. Jhon, and P.-O. Löwdin, Int. J. Quantum Chem., Quantum Biol. Symp. **14**, 189 (1987).

[95] P.-O. Löwdin, Rev. Mod. Phys. **35**, 724 (1963).

[96] E. Clementi, J. Mehl and W. v. Niessen, J. Chem. Phys. **54**, 508 (1971).

[97] J. A. Piccirilli, T. Krauch, S. E. Moroney, and S. A. Benner, Nature **346**, 548 (1990).

[98] C. Switzer, S. Moroney, and S. Benner, J. Am. Chem. Soc. **111**, 8322 (1989).

[99] S. Benner, Nachr. Chem. Tech. Lab. **38**, 442 (1990).

[100] S. Arnott, P. J. C. Smith, and R. Chandrasekaran, in CRC Handbook of Biochemistry and Molecular Biologie (CRC,Cleveland,1976) **Vol. 2**, 411 (1976).

[101] C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).

[102] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).

[103] R. G. Parr and W. Yang, *Density Funktional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

[104] M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A .Peterson, 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. M. Stewart, M. Head-Gordon, C. Gonzalez and J. A. Pople, *Gaussian 94 (Revision D.2)* (Gaussian, Inc., Pittsburgh PA, 1995).

[105] C. A. Chatzidimitriou-Dreismann, Int. J. Quantum Chem. **46**, 483 (1993).

[106] J. T. H. Dunning and P. J. Hay, *Modern Theoretical Chemistry* (H. F. Schaefer, III, New York, 1976), pp. 1–28.

[107] P.-O. Löwdin, Pontificae Acad. Sci. Scr. Varia **31**, 637 (1967).

[108] M. Berry and K. Mount, Rep. Prog. Phys. **35**, 315 (1972).

[109] T. O'Malley, Adv. At. Mol. Phys **7**, 223 (1971).

[110] F. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1959).

[111] R. M. F. Streffer, O. Hübner, and C. A. Chatzidimitriou-Dreismann, in *New Frontiers in Theoretical Biology*, edited by C. A. Dreismann (Hardronic Press, Palm Harbour, 1996).

[112] I. Prigogine, *Form Being to Becoming* (Freemann, San Francisco, 1980).

[113] P.-O. Löwdin, Adv. Quantum. Chem. **19**, 88 (1988).

[114] E. Brändas and N. Elander (eds.), *Resonances – The Unifying Route Towards the Formulation of Dynamical Processes*, Lecture Notes in Physics, Vol. 325 (Springer, Berlin, 1989).

[115] C. A. Chatzidimitriou-Dreisman, Adv. Chem. Phys. **80**, 201 (1991).

[116] J. S. Kwiatkowski, T. J. Zielinski, and R. Rein, Adv. Quantum Chem. **18**, 85 (1986).

[117] V. Hrouda, J. Florián and P. Hobza, J. Phys. Chem. **97**, 1542 (1993).

[118] C. A. Chatzidimitriou-Dreismann, Helv. Chim. Acta **75**, 2252 (1992).

[119] P. G. Jasien and G. Fitzgerald, J. Chem. Phys. **93**, 2554 (1990).

[120] A. N. Isaev, B. Acad. Sci. **40**, 1618 (1991).

[121] R. Czerminski, K. Szczepaniak, W. B. Person, and J. S. Kwiatkowski, J. Mol. Struct. **237**, 151 (1990).

[122] J. Lipinski, Chem. Phys. Lett. **145**, 227 (1988).

[123] T. N. Lively, M. W. Jurema, and G. C. Shields, Int J. Quantum Chem., Quantum Biol. Symp. **21**, 95 (1994).

[124] I. R. Gould and P. A. Kollman, J. Am. Chem. Soc. **116**, 2493 (1994).

[125] R. Santamaria and A. Vazquez, J. Comput. Chem. **15**, 981 (1994).

[126] S. Schreiner and C. W. Kern, J. Am. Chem. Soc. **101**, 4081 (1979).

[127] S. Schreiner and C. W. Kern, Chem. Phys. Lett. **57**, 331 (1978).

[128] V. Hrouda, J. Florian, M. Polasek, and P. Hobza, J. Phys. Chem. **98**, 1457 (1994).

[129] M. L. Bender, R. J. Bergeron, and M. Komiyama, *The Organic Biochemistry of Enzymatic Catalysis* (Wiley, New York, 1984).

[130] J. Sponer, J. Leszczynski, and P. Hobza, J. Bio. Struct. Dyn. **14**, 117 (1996).

[131] J. Florián, V. Hrouda and P. Hobza, J. Am. Chem. Soc. **116**, 1457 (1994).

[132] I. K. Yanson, A. B. Teplitsky and L. F. Sukhodub, Biopolymers **18**, 1149 (1979).

[133] N. C. Seeman, J. M. Rosenberg, F. L. Suddath, J. J. P. Kim and A. Rich, J. Mol. Biol. **104**, 109 (1976).

[134] Q. Zhang, R. Bell, and T. N. Truong, J. Phys. Chem. **99**, 592 (1995).

[135] V. Barone, C. Adamo, and F. Lelj, J. Chem. Phys. **102**, 364 (1995).

[136] L. Meschede and H. Limbach, J. Phys. Chem. **95**, 10267 (1991).

[137] A. Stöckli, B. H. Meier, R. Kreis, and R. R. Ernst, J. Phys. Chem. **93**, 1502 (1990).

[138] S. F. Boys and F. Bernardi, Mol. Phys. **19**, 553 (1970).

[139] P. Hobza and R. Zahradnik, Chem. Rev. **88**, 871 (1988).

[140] J. M. Rosenberg, N. A. Seeman, R. O. Day, and A. Rich, J. Mol. Biol. **104**, 145 (1976).

[141] J. Florián, J. Leszczyński and S. Scheiner, Molec. Phys. **84**, 469 (1995).

[142] R. G. Fowler, G. E. Degnen and E. C. Cox, Mol. Gen. Genet. **133**, 179 (1974).

[143] A. Fersht and Knill-Jones, J. Mol. Biol. **165**, 633 (1983).

[144] M. Topal and J. R. Fresco, Nature **263**, 289 (1976).

[145] D. Brutlag and A. Kornberg, J. Biol. Chem. **247**, 241 (1972).

[146] A. R. Fersht, Trends Biochem. Sci. **5**, 272 (1980).

[147] L. D. Williams, N. G. Williams, and R. B. Shaw, J. Am. Chem. Soc. **112**, 829 (1990).

[148] R. MacPhail, L. D. Williams, D. A. Jones, and R. B. Shaw, J. biomol. Struct. Dynam. **9**, 881 (1992).

[149] F. Kleinjung, *Sensorik und Kinetik biochemischer Nukleinsäure-Wechselwirkungen an Oberflächen* (Ph.D. thesis, Universität Potsdam, 1998).