Celotna doktorska disertacija, vključno s prilogami, naj ne bi obsegala več kot 200 strani.

Vsako glavno poglavje se mora pričeti na lihi strani.

Poglavja si morajo slediti v zaporedju, predpisanem s predlogo.

V kolikor so v delo vključeni algoritmi in/ali slike in/ali tabele, so kazala zanje obvezna.

Pri navajanju virov in literature je potrebno upoštevati pravila, opisana v dokumentu Navajanje virov.doc.

V prilogi je obvezno našteti lastne objave iz disertacije.

Complete doctoral dissertation including appendices should not exceed 200 pages.

Each main chapter should begin on an odd page.

The chapters should follow as prescribed in the template.

When the work contains algorythms and/or figures and/or tables their indexes are obligatory.

The references should consider the rules as described in the document Citation style.doc.

Publications of the candidate from the dissertation must be listed in appendix.

KRATKA NAVODILA
SHORT INSTRUCTION

Jure Pražnikar

Računanje elektrostatskega potenciala okrog proteinov

Doktorska disertacija

Calculating electrostatic potential around proteins

Doctoral Dissertation

Supervisor: Prof. Dr. Dušan Turk

June 2009

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL
Ljubljana, Slovenija

Index

Abstract

X-rays are diffracted on electron clouds and thereby provide a mean to observe 3-dimensional structure of molecules at atomic level. Using medium resolution diffraction data (2Å and lower) we have shown that averaging of electron density maps generated from sets of randomly distributed atoms and can reduce noise and model bias of the maps and thereby assist in their correct interpretation.

When the crystals are so well ordered that diffraction data at ultra-high resolution can be collected (beyond 0.6Å), these data provide information not only about the core electrons surrounding the atomic nucleus, but also contain information about the fine distribution of electron density. Such data thus provide the insight into distribution of electron density clouds in covalent bonds as well as shifts of electron density. To approach interpretation of experimental data to the theoretical quantum mechanical models, we have developed a crystal structure refinement procedure based on mathematical model of quantum mechanics and named it "wave function coefficients" (WFC). WFC electron density is calculated from one "quasi-molecular orbital" which is the sum of all atomic orbitals of the molecule. Our current model, analogously with quantum mechanical models of the Hartee-Fock level, does not consider the effects of atomic movements and distributions encompassed in crystallographic B-factor. The obtained WFC representation of electron density was used to calculate the electrostatic potential around the molecules. To take into account the solvation effects of the media by solving the Poisson Boltzmann equation point charge models must be provided. Partial charges were determined by minimizing the least-squares difference between the electrostatic potential and the potential from the point charges. In the case of protein model unexpectedly large main-chain charge fluctuations were observed thus challenging the current understanding of polarizability of amino acid residues in protein structures.

Povzetek

3-dimenzionalna struktura (makro)molekul je interpretacija merjenih podatkov. V primeru kristaliziranih molekul, gre pri tem za interpretacijo difrakcijskega vzorca kristalne mreže pridobljene s sipanjem rentgenskih žarkov na atomih v kristalu. Na srednji resoluciji difrakcijskih podatkov (2Å in nižje) smo pokazali, da povprečena elektronska gostota izračunana iz serije brcnjenih modelov lahko zmanjša šum in tako omogoča lažjo, ter bolj pravilno interpretacijo elektronske gostote.

Če so molekule v kristalu zelo dobro urejene lahko izmerimo podatke zelo visokih ločljivosti (nad 0.6Å). Pri višjih resolucijah, pa meritve vsebujejo tudi podatke, ki razkrivajo razporeditev gostote ne le za notranje orbitale, temveč tudi sliko gostote elektronskih oblakov kovalentnih in aromatskih vezi - to je območja kemijskih reakcij. Da pa bomo lahko približali interpretacijo eksperimentalnih podatkov s teoretično kvantno mehanskimi računi smo razvili metodo; “koeficienti valovne funkcije” (WFC). WFC elektronska gostota se izračuna iz ene "kvazi-molekularne orbitale", ki je vsota vseh atomarnih orbital v molekuli. Trenutni model, ki je analogen kvantno mehanskemu modelu na Hartee-Fock ravni ne upošteva učinkov termičnega gibanja atomski jeder. Z metodo WFC smo izračunali eksperimentalni eklektrostski potencial. Ker 3-dimenzionalne strukture makromolekul v kristalu določene pri še tako visokih ločljivostih ne vsebujejo pozicij vseh posameznih molekul topila, niti njihove natančne kemijske sestave, je najbolj smiselna rešitev za upodabljanje elektrostatskega potenciala reševanje Poisson-Boltzmannove enačbe. Pisson-Boltzmannov model upošteva, oziroma opisuje razporeditev molekul topila. Parcialne naboje, ki smo jih prilagajali (vsota najmanjših kvadratov) eksperimentalnem elektrostskemu potencialu, smo uporabili za vhodni podatek pri reševanju Poisson-Boltzmannove enačbe. Pri modelu proteina lahko opazimo velike fluktuacije prilagajanih nabojev glavne verige in s tem veliko polarizabilnost aminokislin v proteinskih strukturah.

Abbreviations

RRMS	=	Relative root mean square
CC	=	Linear correlation coefficient
UN	=	Simple uniform weighted map
ML	=	Maximum likelihood weighted map
AK	=	Averaged kick map
GTO	=	Gaussian type orbitals
STO	=	Slater type orbitals
FMO	=	Fragment molecular orbital method
Å	=	Ångström [10^{-10} m]
GOP	=	Gross orbital population
HF	=	Hartree Fock
TZV	=	Triple zeta valence
vdW	=	van der Waals (radius)
FT	=	Fourier transformation
FFT	=	Fast Fourier transformation
WFC	=	Wave function coefficients

1 Introduction

Macromolecular crystallography has been for the last few decades the main source of structural information for macromolecular systems of biological interest. The unique power of macromolecular crystallography is its capacity to obtain a direct image of the electron density for nucleic acids and proteins in an aqueous environment close to that present in biological systems. However, due to the flexibility inherent in macromolecules and to the presence of a significant quantity of water molecules in the crystal, the crystalline order is in general perturbed, and therefore the resulting diffraction is often limited to resolutions between 2 and 3 Å. Since at this resolution range the corresponding electron density maps do not show independent atomic positions, stereochemical information obtained from small molecule studies had to be superimposed to build and refine atomic models. As a result, the stereochemical details of the final models, such as bond distances and angles are strongly biased towards the initially superimposed ones. Another important consequence of the limited resolution in x-ray crystallography was the impossibility to observe hydrogen atoms and therefore of determining protonation states. In the last decade, a series of technical and methodological advances changed this situation and pushed forward the limits of resolution. These advances included, among others, the availability of brilliant and focused synchrotron sources of x-rays, the possibility of performing data collection at cryogenic temperatures and the improvement of crystallographic software ranging from data treatment to refinement. A number of macromolecular structures solved at atomic (1Å) resolution began to appear (Schmidt et al., 2003; Petrova & Podjarny, 2004). At this resolution, atoms are clearly resolved and geometrical restraints (Konnert, 1976; Konnert & Hendrickson, 1980) can be applied with lower weights, leading to accurate measurements of fine details and, sometimes, observation of significant deviations from the standard stereochemical values. Hydrogen atoms start to be visible and protonation states of the polar residues in the active site of the enzyme can be determined (Minasov et al.,2002). Furthermore, observed protonation states can differ from those observed in solution for isolated amino acids. These observed differences may have strong implications for biological functions, and their experimental determination is therefore crucial. In the last few years, the limit of the resolution has been pushed further, and currently a handful of structures are available at subatomic resolutions (better than 0.85 Å). At this resolution level it is possible to obtain information beyond the position of spherical atoms with anisotropic vibrations. In particular, it is possible to obtain residual maps which clearly show valence electrons, including those of hydrogen atoms (Coppens, 1967, 1997, 1998; Lecomte, 1995; Koritsanszky & Coppens, 2001). The consequences of such a direct observation of the electron density from diffraction amplitudes are far-reaching. It should, in theory (Hohenberg & Kohn 1964), be possible to measure the occupation of quantum orbitals, and therefore to determine the reactivity of a given atom, directly from the diffraction data. Such a goal has not yet been attained for macromolecules of biological interest, but a number of efforts are being put forth in this direction (Jayatilaka 1998; Jayatilaka & Grimwood 2001; Bethanis et al 2002).

1.1 Gaussian functions and scattering factors

In crystallography, the atomic form factor, or atomic scattering factor, is a measure of the scattering amplitude of a wave by an isolated atom. These scattering functions are independent of the wavelength of radiation and only depend on the scattering angle and the type of atom. The common feature of all form factors is that they involve a Fourier transform of a spatial density distribution of the scattering object from real space to reciprocal space. To a first approximation, these scattering functions are determined assuming that the electron density of each atom is a discrete and spherically symmetric entity. The spatial density distribution can be expressed as a function of radius, $\rho(r)$, so that the form factor, $f(Q)$ is defined as :

$$f(Q) = \int_0^\infty \rho(r) e^{iQ \cdot r} dr \qquad (1)$$

where $\rho(r)$ is the spatial density of the scatterer about the center of mass of the scatterer ($r = 0$), and Q is the momentum transfer. As a result of the nature of the Fourier transform, the broader the distribution of the

scatterer ρ in real space r, the narrower the distribution of f in Q; i.e., the faster the decay of the form factor. For crystals, atomic form factors are used to calculate the structure factor for a given Bragg peak of a crystal. The X-ray scattering factor is evaluated as the Fourier transform of the electron density distribution of an atom or ion, which is calculated from theoretical wave functions for free atoms.

The incoming X-rays are scattered by the electrons of the (protein) atoms. As the wavelength of the X-rays (1.5 to 0.5 Å) is of the order of the atom diameter, most of the scattering is in the forward direction. It is also obvious that the X-ray scattering power will depend on the number of electrons in the particular atom. The X-ray scattering power of an atom decreases with increasing scattering angle and is higher for heavier atoms. A plot (Figure 1) of scattering factor f in units of electrons vs. $sin(\theta)/\lambda$ shows this behavior. Scattering functions are independent of the wavelength of radiation and only depend on the scattering angle and the type of atom. At zero scattering angle the scattering function of a given atom has a value equal to the number of electrons in the atom.

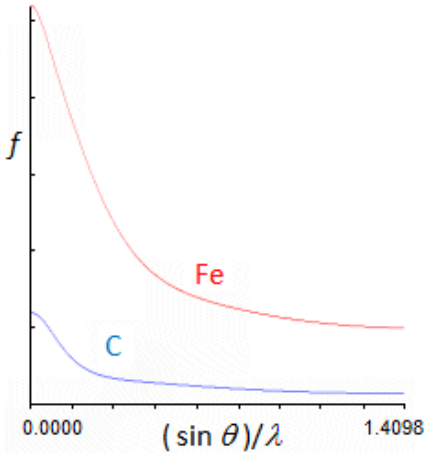


Figure 1 *Atom scattering factors – Fe and C.* Scattering factor of stationary C and Fe atoms plotted as a function of Bragg angle for incident X-ray wavelength of 0.70930 Å. Ticks on the horizontal axis correspond to Bragg angle increments of 10 degrees; ticks on the vertical axis are increments of 5 electrons.

1.2 Cromer Mann coefficients

The need for good values of the atomic scattering factors in crystallographic studies has long been appreciated. It is generally conceded that the best values of the atomic scattering factors are those obtained from the most sophisticated and presumably most realistic atomic models. In particular, the Hartree-Fock values have found particular favor although the difficulty of obtaining true Hartree-Fock wave functions has precluded the compilation of form factors for anything but a limited number of elements. A simplification of the basic Hartree-Fock problem which was introduced (Slater 1951) provides one with a computationally tractable problem. The resulting Hartree-Fock-Slater wave functions are very good approximations to those obtained by Hartree-Fock. Self consistent wave functions have been calculated (Herman & Skillman 1962) for all ground state neutral atoms from Z=2 to Z=103 using the Hartree-Fock-Slater approach.

The normalized scattering curves have been fitted to a 9-parameter equation by Don Cromer and J. Mann (Cromer&Mann 1967). Knowing the 9 coefficients, a_i, b_i and c, and the wavelength, the scattering factor of each atom at any given scattering angle (see equation 2, 3, and 4) can be calculated.

$$f_{analytic}(s) = \sum_{i=1}^{4} a_i e^{-b_i s^2} + c \tag{2}$$

$$s = \sin\theta / \lambda \tag{3}$$

The coefficients along with the maximum and minimum percentage deviations of the analytic fit

defined as:

$$E = \frac{(f_{HF} - f_{analytic})}{f_{HF}} \quad .$$

1.3 Structure factor

Single crystal x-ray diffraction data leads to structure factor amplitudes. The structure factors *F(h)* are the Fourier transforms of the electron density *ρ(r)* of the unit cell of volume V and cell parameters a_i , i = 1, 3:

$$F(h) = \int_{unitcell} \rho(r) \exp(2\pi i h r) d^3 r \tag{5}$$

where $r = xa^*_1 + ya^*_2 + za^*_3$ is a point inside the unit cell (the 'fractional coordinates' *x*, *y* and z have values between 0 and 1) and $h = ha^*_1 + ka^*_2 + la^*_3$ is a point of the 'reciprocal lattice', defined by the unit vectors $a^*_i = (a_j \wedge a_k)V^{-1}$. Since the crystal is triply periodic, the Fourier transform has non-zero values only for integer values of (*h, k, l*), known as the 'Miller indices'. The range of these values is defined as the resolution of the data. The result of the diffraction experiment is the amplitude of the complex quantity $F(h) = F(h) \exp(i\varphi(h))$ and we must know both their amplitude *F(h)* and phase *φ(h)* for calculating *ρ(r)* by inverse Fourier transform. An atomic model can then be superimposed to this density, leading to the map 'interpretation'. The interpretation is facilitated since the electron density is mainly concentrated around atomic positions, and can therefore be expressed as a sum of contributions of individual atoms:

$$\rho(r) = \sum_j \rho_j(r - r_j) \tag{6}$$

where the variable *j* runs over all atoms, and r_j is the position of the *j* th atom. For an ideal crystal, without thermal vibrations and with exactly the same molecules in all unit cells, the electron density ρ_j corresponds strictly to the atomic electronic cloud, which is mostly a sharp Gaussian around the nuclei. However, in practice, there are two important contributions that widen this distribution. One of them is the thermal vibration of atoms, and the second one is the variation in atomic positions in different unit cells. For experiments conducted at a single temperature, these two effects cannot be separated and are therefore grouped in a single Gaussian distribution of the atomic centers. The width of this Gaussian distribution is defined by the value of the mean square displacement of the atomic centers from their mean positions <*u2*>. In real space, the observed electron density is obtained by convoluting this Gaussian distribution with the shape of the atom at rest. By taking the Fourier transform, this convolution is transformed in to a product, leading to the following formula for structure factor:

$$F(h) = \sum_j f_j(|h|) \exp(2\pi i h r_j) \exp(-B_j \frac{h^2}{4}) \tag{7}$$

where r_j are the atomic positions, f_j the radial part of the Fourier transform of the atomic shape of the electron density of the free neutral atom and B_j is the isotropic Debye–Waller factor for each atom. The Debye–Waller factor *B* is related to the mean square displacement as:

$$B = 8\pi^2 \langle u^2 \rangle \quad . \tag{8}$$

The temperature factors or *B*-values, as determined crystallographically, are linearly related to the mean square displacement of an atom and give an indication of atomic flexibility in the crystalline state. The B-factors represent smearing of atomic electron densities around their equilibrium positions due to thermal motion and positional disorder.

1.4 Electron density

The amplitudes and the phases of the diffraction data from the protein crystals are used to calculate electron density map of the repeating unit of the crystal. This map then has to be interpreted as polypeptide chain with a particular amino acid sequence. The interpretation of the electron density map is complicated by several limitations of the data. First of all, the map itself contains errors, mainly due to errors in the phase angles. In addition, the quality of the map depends on the resolution of the diffracted data, which in turn depends on how well-ordered crystals are. This directly influences the image that can be produced.

1.5 Low and middle resolution

Once an electron density map becomes available, atoms may be fitted into the map using computer graphics to give an initial structural model of the protein. A bottleneck in protein crystallography is interpreting an electron-density map; that is, fitting a molecular model to the 3D picture crystallography produces. It is desired to construct a chemically sensible model of the macromolecule, which represents the experimental electron density distribution with a set of labelled atoms and their corresponding coordinates.

Building the initial model is a trial and error process. Use of reliable density maps is crucial for rapid and successful crystal structure determination. After crystallographic phases have been obtained an iterative procedure is used to cycle through density map calculation, molecular model building, rebuilding and refinement until the consistency of the model with the experimentally measured structure factors is maximized. When experimental phases are available, they provide a source of phasing information independent from the model. However, in the molecular replacement case the model is the sole source of phasing information, which, by transformation into density maps, guides model building and rebuilding (Rossmann, 1972). The density maps have the potential to reveal more information than is provided by the current working model. Simultaneously, they are the source of misleading information: it is typically the case that the molecular replacement models used for phasing may be partially incorrect and thus bias the resulting maps. Sometimes a thin line separates the correct interpretation of a density map from the incorrect one. Therefore, it is important to derive maps that are most consistent with the true structure and have the highest potential to identify the incorrectly built parts. In the initial phase, the search model, positioned by the molecular replacement, is used to calculate the electron density map. In the intermediate phase, the models are partially refined and more or less complete. They still contain regions with errors and lack of flexible loops and ligands. The positions of the residues still need to be examined and adjusted to best fit the electron density maps. In the final phase the last remaining, weak density and dubious map features require interpretation. They are commonly occupied by ligands attached to the macromolecular structure or flexible, likely surface located regions and residue side chains, and exhibit larger degree of disorder, when compared to the core of the structure. For solving such kind of problems we have developed method named “kick” maps. (described in methods)

1.6 High resolution – Anisotropic temperature factor

A common practice in X-ray crystallographic structure refinement has been to model atomic displacements or thermal fluctuations as isotropic motions. Recent high-resolution data reveal, however, significant departures from isotropy, described by anisotropic displacement parameters modeled for individual atoms.

The subject of an X-ray diffraction experiment and the subsequent analysis is not a single molecule at an instantaneous point in time. Instead the data collected on a crystal correspond to an ensemble of zillions of molecules observed for a time which is very long compared to typical time-scales of molecular motions. Therefore the result of the experiment, the so-called 'crystal-structure' of a molecule, does not provide a sharp position for each atom but, due to the time and space averages, a three dimensional probability density that is characterized by a mean position and some quantity that is related to the mean displacement of the particular atom from this mean position. Atomic positions and B-factors are adjusted by optimizing the agreement between observed structure factors (F_{obs}) and structure factors calculated on an actual model (F_{calc}). The overall mean displacement of an atom originates from several sources:

- different conformations in different unit cells
- vibration or dynamic transitions within molecules
- lattice defects
- lattice vibrations

From the variety of these contributions it is clear that an isotropic description of mean atomic displacements is only a very crude approximation. In contrast to small molecules the refinement of more detailed models by introducing more parameters into the refinement process is unfortunately not supported by the number and quality of the X-ray data for most macromolecules. The situation is different if atomic resolution data are available. Due to the large number of observables (typically on the order of 30 to 50 refections per non-hydrogen atom) the isotropic model can be upgraded to an anisotropic model (6 parameters to describe the orientation and the elongation of an ellipsoid). In this case, the Debye–Waller factor can be expressed as a symmetric 3×3 matrix B_j, and structure factors are calculated using the formula:

$$F(h)=\sum_j f_j(|h|)\exp(2\pi i h r_j)\exp(-\frac{1}{4}h^T B_j h) \tag{9}$$

When sufficient data are not available restraint should be introduced. Roughly restraints (Konnert 1980) fall in two categories: 1) restraint is based on the fact that a covalent bond between two atoms is fairly rigid so that if the atoms move they will move in phase, 2) restraint is based on the assumption that displacements of atoms that are spatially close will have similar amplitudes and similar directions, this restraint is an extension of the restraints commonly used in isotropic B-factor refinement.

1.7 High resolution – Multipoles

Crystallography at high and ultra-high resolution permits the observation and measurement of the non-spherical character of the atomic electron density. The electron cloud around the atoms of a molecule is deformed as a result of chemical bonding and non-bonding interactions (notably hydrogen bonds) between the atoms. Accurate electron density distribution in the crystalline state can be derived from an ultra-high resolution X-ray diffraction experiment (Coppens, 1967). Such crystal structures can be refined with a Hansen and Coppens (ref) non-spherical multipolar pseudoatom model.

The presence of significant densities in deformation maps at high resolution is an indicator that spherical free atom models, even with anisotropic thermal parameters, give too crude an approximation to the real electron density. The so called ‘multipolar model’ (Hansen and Coppens 1978, Coppens 1997, Spackman 1998) which is widely used in small-molecule crystallography, gives a better representation of electron charge densities. In this model, the total atomic electron density ρ is composed of three parts: (1) ρ_{core}, corresponding to the spherically symmetric core electrons; (2) ρ_{val} , corresponding to the spherical part of the valence electrons and (3) the non spherical part of valence electron densities caused by chemical bonding (Hansen and Coppens 1978). This third part is modeled by a sum of multipolar pseudo-atoms lying in the atomic positions. The valence electron density of such a pseudo-atom is projected on the basis of real spherical harmonics functions $Y_{lm}(\theta,\phi)$ centered on each pseudo-atom. The result is given in the formula:

$$\rho(r)=\rho_{core}(r)+\kappa^3 P_v \rho_{val}(\kappa,r)+\sum_{l=0,l\max}\kappa'^3 R_l(\kappa',r)\sum_{m=\pm 1} P_{lm}Y_{lm}(\theta,\phi) \tag{10}$$

The radial functions $R_l(r)$ used here are of Slater type: $R_l(r) = r^{nl} \exp(-\kappa'\zeta r)$. The number of multipolar parameters P_v, P_{lm}, κ, κ' for each atom increases with the number of included spherical harmonics. Usually, non-hydrogen multipolar atoms have a minimum of 28 parameters (18 multipolar, 3 positional, 6 temperature parameters, 1 occupancy). In cases when the observation/parameter ratio is high enough, parameter values can be obtained directly from least squares refinement against the structure factor amplitudes (Guillot et al 2001). It is possible to molecular properties such as electrostatic potential (Ghermani et al 1993), electric field, net charges, higher moments (Spackman 1992) and topology of the electron density (Souhassou and Blessing 1999). For the case of protein residues, a first approximation to the atomic multipolar atomic charge density parameters can be directly transferred from a pre-calculated database (Pichon–Pesme et al 1995, Jelsch et al 1998). This transfer results in a considerable saving of calculation time, compared with the refinement procedure.

1.8 Quantum mechanics and crystallography

Quantum crystallography concerns the combining of X-ray diffraction data for crystals with quantum mechanics with the the objective of obtaining accurate wave functions that are consistent with the X-ray data (Massa,1995). Quantum mechanics and crystallography are readily combined through the fact that the

electron distributions around atoms are the sources of X-ray scattering and electron density distributions are readily described by quantum mechanical models. The manner in which experiment and theory are connected is through structure factor magnitudes, quantities that are readily obtained from the measured X-ray diffraction intensities, and also obtained from certain Fourier transforms of the quantum mechanical description of the electron density distribution.

The latest advances in obtaining ultrahigh resolution x-ray data for proteins open perspectives for future use of quantum crystallography to determine reactivity of macromolecules. This possibility is based on the Hohenberg–Kohn theorem (Hohenberg and Kohn 1965), together with the Kohn and Sham equations (Kohn and Sham 1965), which established a link between the electron density of a molecule and its wave function. It means that the detailed knowledge of the electron density could give additional information sufficient to accurately determine the wave function. The knowledge of the wave function, in turn, is sufficient to obtain the chemical indices required to understand the reactivity of a molecule. Thus, theoretically, the knowledge of the electron density at ultrahigh resolution may provide all the information required to fully characterize the local subtleties of its chemical nature. The main difficulties with this trend are related to the complexity of mathematical expressions and to the limits in computing power. Therefore, these approaches were applied only to small molecules (Jayatilaka 1998, Grimwood and Jayatilaka 2001, Bytheway et al 2002).

The wavefunction is the fundamental entity that appears in the Schrodinger picture of quantum theory. It is the most compact way to represent all information contained in a system. It is well known that observables are obtained from the wavefunction as expectation values of operators, after an integration over basis states. Thus, an experimental wavefunction can only be determined indirectly using physical observables. The following question now arises: is it possible, even in principle, to determine the wavefunction from the experimental measurement of some observable or set of observations? Hohenberg & Kohn (1964) showed that, within non-relativistic quantum theory, there is a one-to-one mapping between the ground-state electron density - an `observable' from the X-ray experiment - and the exact ground-state wavefunction. Unfortunately, they were not able to give any systematic way to elucidate the nature of this functional relation or how it may be inverted to obtain the wavefunction from the density. Nevertheless, this establishes the density as an important quantity for determining any wavefunction. Even if there was a known procedure to determine the wavefunction from a given electron density, the unavoidable presence of experimental errors means that there is always a range of acceptable wavefunctions which would be compatible with the measurements. In fact, even with perfect data, there would remain the problem that the electron density is a continuous function, requiring an infinite number of measurements to define. The electron density is not, strictly, an observable. In fact, the electron density emerges as an `observable' related to the measured X-ray scattering data only after a detailed analysis of the Schrodinger equation for the entire crystal in which many approximations have been made - including, for example, the Born-Oppenheimer approximation, the first Born approximation, harmonic approximation for the nuclear potential and the rigid pseudoatom approximation for the electron density (Stewart, 1977; Stewart & Feil, 1980). Only with these approximations in mind, can it be said that the elastic X-ray diffraction experiment on a crystal yields the magnitude set of structure factors *F(h)*, which are related to the electron density *ρ(r)* in the unit cell (averaged over the thermal motions of the nuclei) by the Fourier transform,

$$F(h) = \int_{cell} \rho(r) \exp(2\pi i \, \mathbf{r} \cdot \mathbf{B}\mathbf{h}) \, d\,\mathbf{r} \tag{11}$$

where B is the reciprocal-lattice matrix, depending only on the shape of the crystal unit cell and $h=[h_1, h_2, h_3]$ are integers called the Miller indices which describe the direction of the scattered X-ray beam.

1.9 Partial point charges

Partial charges are created due to the asymmetric distribution of electrons in chemical bonds. The resulting partial charges are a property only of zones within the distribution, and not the assemblage as a whole. For example, chemists often choose to look at a small space surrounding the nucleus of an atom: When an electrically neutral atom bonds chemically to another neutral atom that is more electronegative, its electrons are partially drawn away. This leaves the region about that atom's nucleus with a partial positive charge, and it creates a partial negative charge on the atom to which it is bonded.

Atomic point-charges are often used to provide simple models for the electrostatic potential of a molecule. These models are widely used in applications such as molecular dynamics simulations and

rationalizing bonding in weakly bound systems. Although atomic charges in a molecule cannot be measured experimentally, they have been used widely for conformational analysis and for interpreting molecular reactivity. For three decades, molecular force fields, such as, CHARMM (Brooks 1983, 2009) , AMBER (Kollman *et al* 2008), and others, have been used successfully for computing structures, and thermodynamic and kinetic properties of proteins, as well as their interactions. Atomic charges in a molecule have been calculated by using quantum mechanical and empirical methods. The atomic charges of the CHARMM, field is empirical, and has been parametrized to reproduce thermodynamic properties of organic liquids and/or crystal structures of organic compounds and peptides, whereas Mulliken charges and ESP-fitted (Singh & Kollman 1984, Bayly 1993, Cornell 1993, Cieplak 1995, Woods & Chappelle 2000, Anisimov 2005) charges were employed for the AMBER force field.

Electrostatic interactions, calculated from atomic charges, have been recognized as one of the essential factors for describing properties of macromolecules. Atomic point-charges are often used to provide simple models for the electrostatic potential. These models are widely used in applications such as molecular dynamics simulations and rationalizing bonding in weakly bound systems. The concept of atomic charges is fundamental to all biochemistry, furthermore atomic charges are enormously powerful in understanding chemical reactivity and physical properties.

1.10 Force field

The paper of Engh & Huber (1991), with its description of accurate geometrical parameters of amino-acid residues, has provided a foundation for the use of geometrical restraints in the refinement of protein structures. A similar step forward in the area of nucleic acids was made by Parkinson et al. (1996). However, these efforts have no counterpart for so-called 'hetero' compounds. As a result, the structures of small molecules found in complexes with biomacromolecules are often less reliable than those of the surrounding amino acids and nucleic acid bases. The reason is most probably the essentially boundless structural diversity of small molecules as opposed to the limited numbers of building blocks of proteins and nucleic acids (Kleywegt et al., 2003). These small molecules are physiological ligands, cofactors, lead compounds, substrate analogs etc. and the accuracy of their structures can be of crucial importance for interpreting their (potential) biological roles. The number and variety of macromolecular structures in complex with 'hetero' ligands is growing. The need for rapid delivery of correct geometric parameters for their refinement, which is often crucial for understanding the biological relevance of the structure, is growing correspondingly. The current standard for describing protein structures is the Engh– Huber parameter set. It is an expert data set resulting from selection and analysis of the crystal structures gathered in the Cambridge Structural Database (CSD). Clearly, such a manual approach cannot be applied to the vast and ever-growing number of chemical compounds. Therefore, a database, named PURY (Andrejašič *et al.* 2009), of geometric parameters of chemical compounds has been developed, together with a server that accesses it.

Following is the minimal equation representing the potential energy, V, of a systems as a function of the structure, r:

$$V(r) = \sum_{bonds} K_b(b-b_0) + \sum_{angles} K_\theta(\theta-\theta_0) + \sum_{dihedrals} K_\chi(1+\cos(n\chi-\delta)) + \sum_{nonbonded-pairs\, i,\, j} \left[\frac{q_i q_j}{4\pi\epsilon_0 r} - \varepsilon_{ij} \left\{ \left(\frac{R_{minij}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{minij}}{r_{ij}} \right)^{6} \right\} \right] \quad (12)$$

Equation may be separated into the internal terms, including the bond, angle, and dihedral contributions, and the non-bonded or external terms that include the electrostatic and van der Waals terms. For the internal terms, b_0 and θ_0 represent the equilibrium bond and angle terms, K_b, K_θ, and K_χ represent the bond, angle, and dihedral force constants, n represents the multiplicity or periodicity of the dihedral term, δ represents the phase of the dihedral term, and b , and θ are the bond, angle, and dihedral angle, respectively, that define r. The dihedral term is a Fourier series allowing for a accurate description of torsional surfaces. Additional terms are often included in the internal portion of the force field. Improper dihedral angles, based on a harmonic term, are used to treat out-of-plane bending motions (e.g., out-of-plane motions of aromatic hydrogens). Concerning the external portion of the force field, q_i and q_j are the partial atomic charges on atoms i and j, ϵ_0 is the dielectric constant, r_{ij} is the distance between atom i and j, and ε_{ij} and R_{minij} are the well depth and minimum interaction distance describing the vdW interaction between atoms i and j.

Bond polarization is not mimicked in most molecular modeling trials. Force fields, such as CHARMM or AMBER, which are used for the refinement of protein structures, have constant partial atomic charges irrespective of secondary structures. Recent progress in methods and in computer facilities have led several groups to develop fast semiempirical quantum mechanical procedures for calculating partial atomic charges of proteins (Liu 2001, Swinnen 2002). One of these procedures was used to calculate the static and dynamic structures of Crambin. The results suggested significant fluctuations of partial charges of main-chain atoms, especially in helices.

1.11 Electrostatic potential

Electrostatics influences various aspects of nearly all biochemical reactions (Baker et al. 2001), such as macromolecular folding and conformational stability. It also determines the structural and functional properties of biological samples, such as shapes, binding energies, and association rates (Sharp and Honig 1990). So the modeling of electrostatics in molecular modeling packages (Humphrey et al. 1996) has great practical, as well as, theoretical importance, for structure-based drug design and protein folding. Electrostatic properties of biological samples can be modeled by quantum mechanical methods or classical electrostatics. Quantum mechanical methods are more accurate, but due to their immense computational demands, can only be applied to small molecules. Classical electrostatics interactions are modeled as interactions between partial atomic charges. They depend not only on 3D structure of molecules and charge distributions, but also on the environment. Biological processes occur in aqueous solution, so solvent plays an important role in determining the electrostatics of solute molecules. Due to the huge computational cost, solvent properties are normally described in terms of average values. That means, instead of treating each solvent atom explicitly, treat them as a continuum with average properties. The more important solute molecules are described in atomic details (Honig and Nicholls 1995). These result in a Poisson-Boltzmann equation for describing the electrostatic interactions in solution.

For solving Poisson-Boltzmann equation charge distribution is need. The charge density $\rho(r)$ present interior charge distribution of the fixed positions of all charges in the molecule and an exterior charge density, which accounts for the redistribution of ions in solution in response to the electric potential in solution. Use of reliable point charges is very important for macromolecular electrostatics.

2 Aims and hypothesis

2.1 Aims

The first aim was to validate and further develop current approaches of electron density map calculations at low and middle resolutions. We will investigated the unweighed and maximum likelihood weighted electron density map calculation and extend their uses by the averaged kick map approach, where each kick map is calculated from atomic coordinates modified by random shifts.

The second aim was to approach the interpretation of X-ray diffraction data with the theoretical to quantum mechanical model. The number of reflections collected at ultra hight resolution provides sufficient number of observables to which wave function coefficients can be fitted directly.

The third aim was to provide experimental basis for calculation of electrostatic potential around molecules and to grasp the potential and possible consequences of this kind of calculation of electrostatic potential around macromolecular models.

2.2 Hypothesis

Introducing noise followed by averaging will improve electron density map quality at lower and intermediate resolutions.

Replacement of standard spherical scattering factors (Cromer&Mann 1967) by mathematical model based on quantum mechanics will provide improved view into structure and electrostatic properties of observed molecules.

3 Methods

3.1 Improvement of map quality

The kick maps are based on structural models with randomly displaced atoms. Random shifts of coordinates, termed kicking, was introduced into the program MAIN (Turk, 1992; Gunčar *et al.*, 2000; Turk 2007) and applied in map calculations and refinement. Each kick modifies atomic coordinates by a random shift within a given interval by which each point within a cube around the atomic position has an equal probability for the new position. From the "kicked" atomic model, structure factors are calculated and used for the difference map calculation, mostly the $2F_{obs}–F_{calc}$ type map. After each map calculation the model coordinates are restored to their original values. The final AK map is the average of the whole series of the kick maps, where F_{model} for each map is calculated from the kicked atomic positions generated with a different random number seed. Each time the model is kicked, the starting seed is changed to produce a novel series of random numbers, unique for each consecutive model modification.

There are two methods for AK map calculation; one can average the F_{model} structure factors and calculate the map using averaged F_{model} or average the individual kick maps. The principal difference between the two is following: in the first case scaling of F_{model} to F_{obs} is performed once for the set of averaged F_{model} and the $2F_{obs}$-F_{model} map is computed, whereas in the second case each individual F_{model} from the series is scaled to Fobs and individual $2F_{obs}$-F_{model} map is computed and added to the sum of maps and averaged at the end. The later approach was found to be superior and is used throughout this work (data not shown).

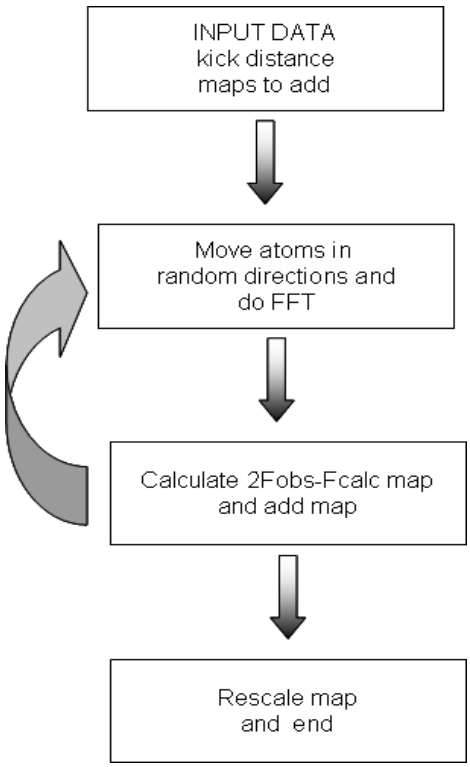


Figure 2 *Kick map algorithm.* Diagram describing the general procedure for calculation of kick map in program MAIN (Turk, 1992).

3.1.1 Electron density map

We used the following maps for all tests discussed below:

- $|F_{model}|exp(i\varphi model)$ map (F_{model} maps, in what follows)
- simple uniform weighted maps $(2|F_{obs}|-|F_{model}|)exp(i\varphi model)$, (UN; $2F_{obs}–F_{model}$)
- maximum-likelihood weighted $(2m|F_{obs}|–D|F_{model}|)exp(i\varphi model)$, (ML; $2F_{obs}–F_{model}$)
- UN Averaged Kick maps (AK)

F_{model} is the total model structure factor that accounts for all atoms and bulk-solvent contribution, as well as various scales (as defined, for example, in Afonine et al., 2005b). The final Fmodel maps generated from the best available structure were used as the reference. The coefficients of the ML maps (Read, 1986, Pannu & Read, 1996, Murshdov et al., 1997) were computed as described in (Lunin & Skovoroda, 1995; Urzhumtsev et al., 1996).

3.1.2 Application

To compare the maps, we have used local map correlation coefficients (CC) and density at positions of the atomic model. For the CC of the local map the region around a part of the atomic model has been selected. A CC greater than 0.8 is generally described as a strong correlation, whereas a CC less than 0.5 is a weak correlation (Lunin & Woolfson, 1993 , Lunin & Skovoroda, 1995).

Table 1 *Ammodytin L*. Data and model statistics used in tests.

PDB	Resol.(Å)	Number of atoms	Data completeness(%)	Number of reflection	R-factor(work/free)
3Dih	2.6	1007	96	5323	0.15 / 0.22

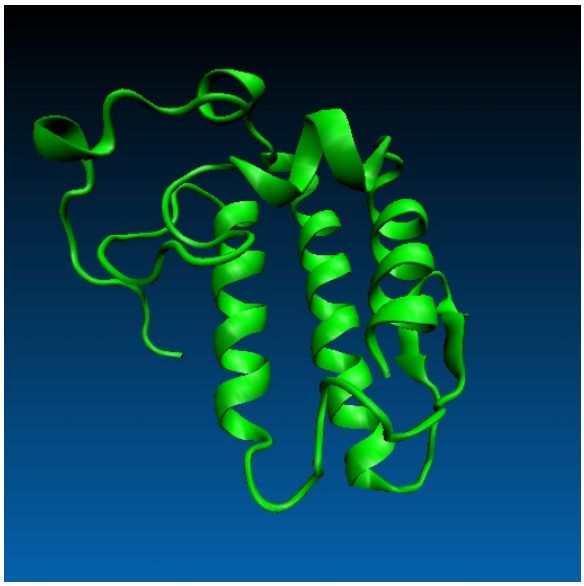


Figure 3 *Structure of Ammodytin L (secondary structure)*. Figure was made with VMD-openGL software. (Humphrey 1996).

3.2 Theoretical calculations

Quantum mechanics is a theory to explain and predict the behavior of particles such as electrons, protons, neutrons, atomic nuclei, atoms, and molecules, as well as the photon-the particle associated with electromagnetic radiation or light. Quantum mechanics is actually the correct mathematical description of the behavior of electrons and thus of chemistry. In theory, quantum mechanic can predict any property of an individual atom or molecule exactly. In practice, the quantum mechanics equations have only been solved

exactly for one electron systems. A myriad collection of methods has been developed for approximating the solution for multiple electron systems. These approximations can be very useful, but this requires an amount of sophistication on the part of the researcher to know when each approximation is valid and how accurate the results are likely to be.

In quantum chemistry each particle is described by a wave function $\Psi(r, t)$, where $\Psi^*\Psi$ represents the probability of finding the particle at position r at time t. To each classical observable, such as position and momentum, there exists a corresponding quantum mechanical operator, which when acting on the wave function, yields the expectation value of this operator. The most important operator, the Hamiltonian, was introduced by Schrödinger and enables to calculate the energy of a given system. The Schrödinger equation in its time-independent form reads

$$\hat{H}\,\Psi = E\Psi \tag{13}$$

where Ψ is a wave function, E the energy and is the Hamiltonian operator:

$$\hat{H} = T + V_{ext} + V_{ee} \tag{14}$$

T is the kinetic energy operator, V_{ext} is the external potential produced by the nuclei acting on the electrons and V_{ee} is the electron-electron potential. In order to obtain a physically relevant solution of the Schrödinger equation, the wave function must be continuous, single-valued, normalizable, and antisymmetric with respect to the interchange of electrons.

The Hamiltonian operator is, in general,

$$\hat{H} = -\sum_{i}^{particles} \frac{\nabla_i^2}{2\mathrm{m}_i} + \sum_{i<j}^{particles} \sum \frac{q_i q_j}{r_{ij}} \tag{15}$$

$$\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \tag{16}$$

where ∇_i^2 is the Laplacian operator acting on particle i. Particles are both electrons and nuclei. The symbols mi and q_i are the mass and charge of particle i, and r_{ij} is the distance between particles. The first term gives the kinetic energy of the particle within a wave formulation. The second term is the energy due to Coulombic attraction or repulsion of particles. This formulation is the time-independent, non-relativistic Schrödinger equation.

In the language of mathematics, an equation of this form is called an eigen equation. Ψ is then called the eigenfunction and E an eigenvalue. The wave function Ψ is a function of the electron and nuclear positions. As the name implies, this is the description of an electron as a wave. It is a probabilistic description of electron behavior. As such, it can describe the probability of electrons being at certain locations, but it cannot predict exactly where electrons are located. The wave function is also called a probability amplitude because it is the square of the wave function that yields probabilities. This is the only rigorously correct meaning of a wave function.

3.2.1 The variational principle

In practice, the wave function of an interacting many-body system is not known a priori. To solve the Schrödinger equation one starts from a trial wave function Ψ_T , which is usually a linear combination of basis functions:

$$\Psi_T = \sum_n c_n \psi_n \tag{17}$$

where ψ_n are orthonormal and normalized wave functions, such that

$$\sum_n |c_n|^2 = 1 \tag{18}$$

An approximation to the true ground-state wave function is found by applying the variational principle, which states: The expectation value of a Hamiltonian, $\hat{H}$, calculated with a trial wave function Ψ_T, is never lower in value than the true ground state energy, ε_0, value calculated with the true ground-state wave function Ψ_0. Minimizing ε_0 with respect to Ψ_n, means basically optimizing the linear combination of coefficients.

3.2.2 Hartree Fock approximation

The computations that are derived directly from theoretical principles with no inclusion of experimental data are called ab initio, the term is Latin for ”from the beginning”. This is an approximate quantum mechanical calculation. The approximations made are usually mathematical approximations, such as using a simpler functional form for a function or finding an approximate solution to a differential equation.

The most common type of ab initio calculation is called a Hartree-Fock calculation, in which the primary approximation is the central field approximation. Which takes into account Coulombic electron-electron repulsion effect. This gives the average effect of the repulsion, but not the explicit repulsion interaction. This is a variational calculation, meaning that the approximate energies calculated are all equal to or greater than the exact energy. The energies are calculated in units called Hartrees (1 Hartree = 27.2116 eV). Because of the central field approximation, the energies from HF calculations are always higher than the exact energy and tend to a limiting value called the Hartree-Fock limit as the basis set is improved. One of the advantages of this method is that it breaks the many-electron Schrödinger equation into many simpler one-electron equations. Each one electron equation is solved to yield a single-electron wave function, called an orbital, and energy, called an orbital energy. The orbital describes the behavior of an electron in the net field of all other electrons. The second approximation in HF calculations is that the wave function must be described by mathematical function, which is known exactly for only a few one-electron systems. The functions used most often are linear combinations of Gaussian-type orbitals [$e^{(-ar2)}$], abbreviated GTO. The wave function is formed from linear combinations of atomic orbitals or, stated more correctly, from linear combinations of basis functions.

Single molecule calculations were performed at the HF-level using GAMESS program package with the molecular geometry from the x-ray model refinement. In the isolated molecule the standard molecular TZV basic set was used for Thym and Heps molecules.

3.2.3 Fragment molecular orbital theory

The fragment molecular orbital method (FMO) was proposed by K. Kitaura and coworkers in 1999 (Kitura 1999). FMO is deeply interconnected with the well known energy decomposition analysis (Kitaura & Morokuma 1976). The main use of FMO is to compute very large molecular systems by dividing them into fragments and performing ab initio quantum-mechanical calculations of fragments and their dimers, whereby the Coulomb field from the whole system is included. The latter feature allows fragment calculations without unphysical caps and results in the proper electron density treatment of the whole system. Since very large systems can be computed with high level ab initio wave functions, FMO is thought to be very useful for biological applications to compute the whole proteins and their complexes.

3.2.4 Mulliken population analysis

Mulliken charges arise from the Mulliken population analysis (Mulliken 1955) and provide a means of estimating partial atomic charges from calculations carried out by the methods of computational chemistry, particularly those based on the linear combination of atomic orbitals molecular orbital method. If the coefficients of the basis functions in the molecular orbital are $C_{\mu i}$ for the $\mu'th$ basis function in the $i'th$ molecular orbital, the density matrix terms are:

$$D_{\mu\nu} = 2\sum_i C_{\mu i} \cdot C_{\nu i}^{\mathbb{C}} \tag{19}$$

or a closed shell system where each molecular orbital is doubly occupied. The population matrix P then has terms

$$P_{\mu\nu} = D_{\mu\nu} S_{\mu\nu} \tag{20}$$

S is the overlap matrix of the basis functions. The sum of all terms of $P_{\mu\nu}$ is N - the total number of electrons. The Mulliken population analysis aims first to divide N among all the basis functions. This is done by taking the diagonal element of $P_{\mu\nu}$ and then dividing the off-diagonal elements equally between the two appropriate basis functions. Since the off-diagonal terms include an $P_{\mu\nu}$ and $P_{\nu\mu}$, this simplifies to just the sum of a row. This defines the gross orbital population (GOP) as

$$GOP_{\mu} = \sum_{\nu} P_{\mu\nu} \tag{21}$$

The GOP_{μ} terms sum to N and thus divide the total number of electrons between the basis functions. It remains to sum these terms over all basis functions on a given atom A to give the gross atom population (GAP). The sum of GOP_A terms is also N. The charge Q_A, , is then defined as the difference between the number of electrons on the isolated free atom, which is the atomic number, Z_A and the gross atom population:

$$Q_A = Z_A - GAP_A \tag{22}$$

For all presented models the theoretical Mulliken charges were obtained directly from the GAMESS.

3.2.5 Basis set

A basis set is a set of functions used to describe the shape of the orbitals in an atom. Molecular orbitals and entire wave functions are created by taking linear combinations of basis functions and angular functions. Most semiempirical methods use a predefined basis set. When ab initio or density functional theory calculations are done, a basis set must be specified. Although it is possible to create a basis set from scratch, most calculations are done using existing basis sets. The type of calculation performed and basis set chosen are the two biggest factors in determining the accuracy of results.

The individual references for Basis set, Contraction shemes and Basis functions have not been included here. Most can be readily found in the articles. (Leland 1960, Weeks 1969, Mitchell 1969, Dunning 1977, Wilson 1983, Huzinaga 1985, Davidson 1986, Feller1990, Almloé 1991, Sordo1992, Thomas 1993, Wilson 1997, Dunning 1998)

Most calculations today are done by choosing an existing segmented GTO basis set. These basis sets are identified by one of a number of notation schemes. These abbreviations are often used as the designator for the basis set in the input to *ab initio* computational chemistry programs. The following is a look at the notation for identifying some commonly available contracted GTO basis sets. The smallest basis sets are called minimal basis sets. The most popular minimal basis set is the STO-3G set. This notation indicates that the basis set approximates the shape of a STO orbital by using a single contraction of three GTO orbitals. Commonly referred to as the Pope basis sets, are indicated by the notation 6-31G. This notation means that each core orbital is described by a single contraction of six GTO primitives and each valence.

3.2.6 Basis functions

The tight functions in the basis set are those having large Gaussian exponents. These functions described the shape of the electron density near the nucleus; they are responsible for a very large amount of total energy due to the high kinetic and potential energy of electrons near the nucleus. However, the tight functions have very little effect on how well the calculation describes chemical bonding. Altering the tight basis functions may result in slightly shifting the atomic size. Although it is, of course, possible to add additional tight functions to an existing basis, this is very seldom done because it is difficult to do correctly and it makes very little difference in the computed chemical properties. It is advisable to completely switch basis sets if the description of the core region is of concern. Diffuse functions are those functions with small Gaussian exponents, thus describing the wave function far from the nucleus. It is common to add additional diffuse functions to a basis. Polarization functions are functions of a higher angular momentum than the occupied orbitals, such as adding d orbitals to carbon or f orbitals to iron. These orbitals help the wave function better span the function space. This results in little additional energy, but more accurate geometries and vibrational frequencies.

3.2.7 Contraction shemes

The orbitals almost always have the functional form given

$$\phi = Y_{lm} \sum_i C_i \sum_j C_{ij} e^{\xi_{ij} r^2} \tag{23}$$

The Y_{lm} function gives the orbital the correct symmetry *(s, p, d, etc.)*. *exp(-r²)* is called a Gaussian primitive function. The contraction coefficients C_{ij} and exponents ζ_{ij} are read from a database of standard functions and do not change over the course of the calculation. This predefined set of coefficients and exponents is called a basis set. An enormous amount of work is involved in optimizing a basis set to obtain a good description of an individual atom. By using such a predefined basis set, the program must only optimize the molecular orbital coefficients C_i.

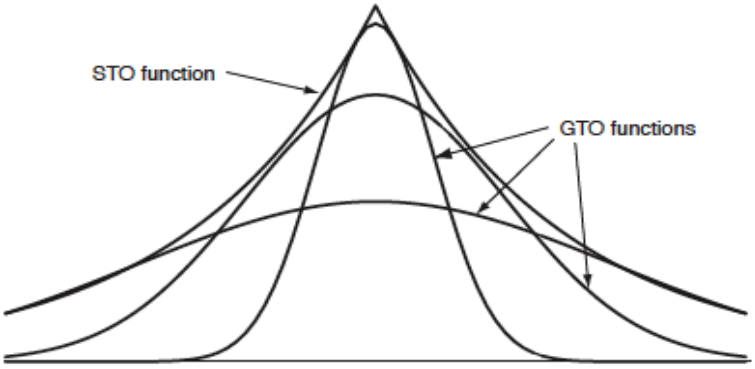


Figure 4 *GTO and STO orbitals.* All of the one-electron orbitals can be built by combining sets of gaussian functions (gaussian primitives) that approximate each STO. The result is called a contracted gaussian function.

The orbitals are referred to as Gaussian type orbitals, or GTO, since they incorporate Gaussian functions, *exp(-ζr2)*. The exact solution to the Schrödinger equation for the hydrogen atom is a Slater type orbital, or STO, of the form *exp(-ζr)*. GTO basis sets require more primitives to describe the wave function than are needed for STO calculations. However, the integrals over GTO primitives can be computed analytically, which is so much faster than the numeric integrals over STO functions that any given accuracy can be obtained most quickly using GTO functions. As such, STO basis sets are sometimes used for high-accuracy work, but most calculations are now done with GTO basis sets.

3.2.8 Application

Crambin (PDB code: 1EJG)

- Resolution: 0.54 Å
- Number of residues: 46
- Solvent content: 30%
- Data were processed with DENZO (Otwinowski and Minor 1997) and DREAR (Blessing et al 1998)
- Refinement was done by Shelxl-97 and Mopro
- R-factor (final): 0.09

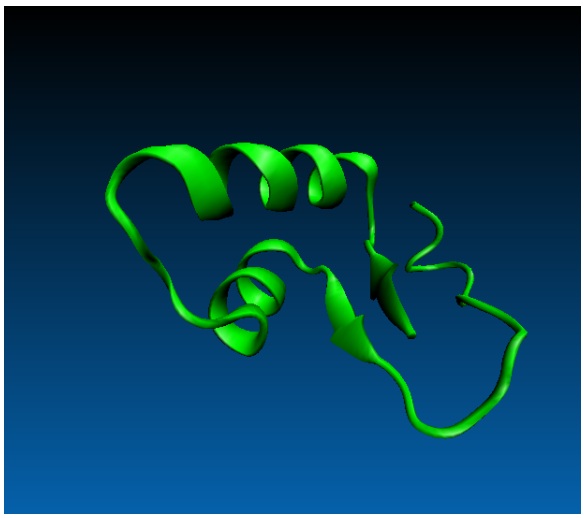


Figure 5 *Structure of Crambin (secondary structure).* Figure was made with VMD-openGL software.

Ab initio calculations for Crambin model were performed at HF/3-21G level using Fragment molecular orbital method. Protein model Crambin has multiple conformations. In all performed calculations only conformation "A" (highest occupancy) was used. Input file for GAMESS was made by FMOutil (Fedorov et al. 2007) program. The exact positions of all hydrogen nuclei are not known, therefore missing hydrogen atoms were added using FMOutil program.

3.3 Wave function coefficient and refinement

Least squares is the simplest statistical method used in macromolecular refinement. Like empirical energy, the history of least squares in macromolecular structure determination extends back to the 1970s (Konnert, 1976) and the approach continues to be used today. The least-squares residual function is defined as:

$$f(p) = \sum_{i}^{all\ data} \frac{\left[Q_o(i) - Q_c(i,p)\right]^2}{\sigma_0(i)^2} \tag{24}$$

where $Q_o(i)$ and $\sigma_o(i)$ are the value and standard deviation for observation number i. $Q_c(i, p)$ is the model's prediction for observation i using the set of model parameters p. The larger the difference between the observation and the model's prediction, the worse the model. The more precisely we know an observation, the more important that observation becomes in the overall sum. One varies the parameters of the model to find a set that gives the lowest sum of deviants. The values of the parameters found by minimizing this function are those that have the smallest individual standard deviation or the smallest probable error (Mandel, 1984).

3.3.1 Refinement and FFT

The way of minimizing a function $f(p)$ which is of most practical importance in modeling biological systems is to solve the problem iteratively. Given a starting point and the derivative at that point produce a new point that is closer to the desired minimum using

$$p_{new} = p_{old} - \eta \frac{df(p)}{dx} \tag{25}$$

The parameter η controls the size of the change. A very time consuming step in the refinement procedure is the calculation of all structure factors and their derivatives from the new parameter set at the end of each cycle in the refinement. To speed up these calculations fast Fourier algorithm are used (FFT). A detailed discussion of the application of the fast Fourirer technique in crystallography is given by Ten Eycke (1973); see also Agarwal (1978). The electron density is sampled at grid points (real space), usually

at distances equal to one-third of the maximum resolution. Equation (26) tells us how to calculate electron density $\rho(x,y,z)$; simply construct a Fourirer series using the structure factors F_{hkl}.

$$\rho(\mathrm{x,y,z}) = \frac{1}{V}\sum_{\mathrm{hkl}} |F_{\mathrm{hkl}}| e^{-2\pi i(\mathrm{hx+ky+lz-}\Phi(\mathrm{h,k,l}))} \tag{26}$$

3.3.2 Application

During positional and wave function coefficients refinement no restrictions were applied. A kick method has been used to overcome the multiple minima problem inherent in finding the global minimum. Each kick modifies atomic coordinates *(X, Y and Z)* for a random shift within a given interval:

```
do i = 1, natoms
    X(i)_{mod} = X(i)_{orig} + KICK*2*(rnd - 0.5))
    Y(i)_{mod} = Y(i)_{orig} + KICK*2*(rnd - 0.5))
    Z(i)_{mod} = Z(i)_{orig} + KICK*2*(rnd - 0.5))
end do
```

by which each point within the cube volume around the original atomic position has equal probability for the new position. Refinement starts from the "kicked" atomic model. The same method was used when wave function coefficients were “kicked”.

R-factor comparison between standard refinement and wave function coefficients refinement was made. Structures using wave function coefficients were refined with MAIN (Turk 1992) program were at several resolutions.

3.4 Implementation and use of Wave function coefficient in MAIN

The electron density calculated using wave function coefficients is expressed as follows:

$$\rho(r) = \left(\Psi(r)\right)^2 \tag{27}$$

where Ψ is defined as:

$$\Psi = \sum_n c_n \phi_n \tag{28}$$

Function Ψ is a sum of atomic orbitals ϕ_n (see equation 28). The number n represents a number of all atomic orbitals. Number n depends on type of basic set and on number of atoms in molecule. Wave function coefficients c_n values can be any real number. No restrictions were introduced.

Following expressions were used for wave function coefficients refinement:

$$\frac{\partial \rho}{\partial c_i} = \frac{\partial \Psi^2}{\partial c_i} = \frac{\partial\left(\sum_i c_i \phi_i\right)^2}{\partial c_i} = 2\Psi\sum_i \phi_i \tag{29}$$

and expression for positional refinement:

$$\frac{\partial \rho}{\partial r} = \frac{\partial \Psi^2}{\partial r} = \frac{\partial\left(\sum_i c_i \phi_i\right)^2}{\partial r} = 2\Psi\sum_i c_i \frac{\partial \phi_i}{\partial r} \tag{30}$$

In case of TZV basic set the oxygen, carbon and nitrogen atom are described with the fourteen functions each, of which five are S-type and nine functions are of P-type, suitable for describing the asymmetric part of the electron density.

3.4.1 Application

Thym
- Resolution: 0.38 Å
- Number of atoms: 31
- Data were processed with DENZO (Otwinowski and Minor 1997)
- Refinement was done by Shelxl-97
- R-factor (final): 0.06

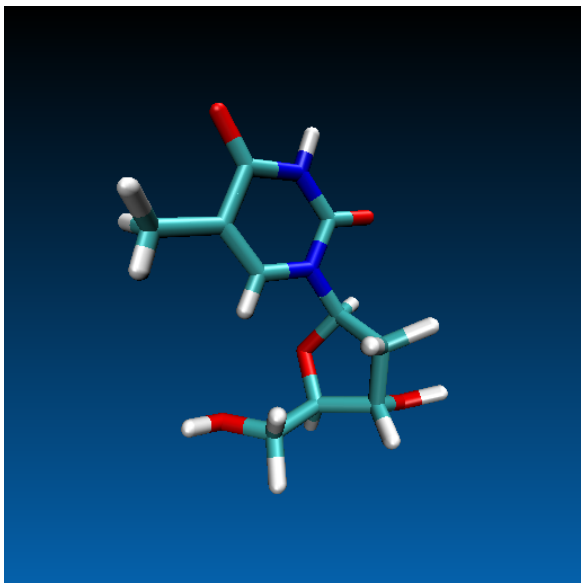


Figure 6 *Thym model.* Figure was made with VMD-openGL software.

Heps
- Resolution: 0.50 Å
- Number of atoms: 33
- Data were processed with DENZO (Otwinowski and Minor 1997)
- Refinement was done by Shelxl-97
- R-factor (final): 0.04

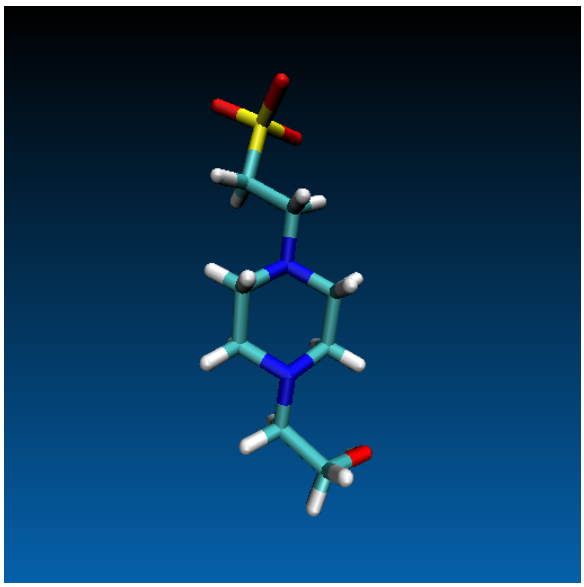


Figure 7 *Heps model.* Figure was made with VMD-openGL software.

Thym and Heps models were provided by Wladek Minor and Maksymilian Chruszcz from Department of Molecular Physiology and Biological Physics, University of Virginia (personal communication).

3.4.2 Example for water molecule

We show simple wave function coefficients for water molecule (H_2O). Equation (31) has 20 wave function coefficients. Wave function coefficients c_1, c_2, c_3, describing electron density around first hydrogen atom, coefficients from c_4 to c_{17} describing electron density around oxygen atom and coefficients c_{18}, c_{19}, c_{20} describing electron destiny around second hydrogen atom.

$$\begin{aligned}\phi = & c_1\varphi(1\mathrm{S}) + c_2\varphi(2\mathrm{S}) + c_3\varphi(3\mathrm{S}) + c_4\varphi(1\mathrm{S}) + c_5\varphi(2\mathrm{S}) \\ & + c_6\varphi(3\mathrm{S}) + c_7\varphi(4\mathrm{S}) + c_8\varphi(5\mathrm{S}) + c_9\varphi(6\mathrm{px}) + c_{10}\varphi(6\mathrm{py}) \\ & + c_{11}\varphi(6\mathrm{pz}) + c_{12}\varphi(7\mathrm{px}) + c_{13}\varphi(7\mathrm{py}) + c_{14}\varphi(7\mathrm{pz}) + c_{15}\varphi(8\mathrm{px}) \\ & + c_{16}\varphi(8\mathrm{py}) + c_{17}\varphi(8\mathrm{pz}) + c_{18}\varphi(1\mathrm{S}) + c_{19}\varphi(2\mathrm{S}) + c_{20}\varphi(3\mathrm{S})\end{aligned} \tag{31}$$

According to equation (27) and (28) the electron density value in each grid is calculated. For illustration a part of source code written in fortran77 language is showed (see Appendix A).

3.5 Charge fitting

The basic idea with electrostatic potential fit charges is that a least squares fitting algorithm is used to derive a set of atom-centered point charges which best reproduce the electrostatic potential of the molecule. The partial charges obtained in this manner may be used as an alternative to the Mulliken and Löwdinn analyses (Lowdin 1970), and as input to force-field programs requiring partial-charges. The analytical electrostatic potential at position r, $V_t(r)$, is defined as:

$$V_t(r) = \int \frac{\rho(r')}{|r - r'|} dr' + \sum_A \frac{Z_A}{|r - R_A|} \tag{32}$$

where $\rho(r')$ is the electronic charge density at point r', Z_A is the nuclear charge of atom A. The sum runs over all atoms, and the integral runs over all space. The `best' partial charges are determined by minimizing the least-squares difference between the electrostatic potential and the potential from the point charges V_q. The latter is simply:

$$V_q(r) = \sum_A \frac{q_A}{|r - R_A|} \tag{33}$$

where q_a is the partial charge assigned to atom A. However, it is also necessary to impose the condition that the partial charges sum up to the total charge on the molecule, Q (Singh & Kollman 1984, Bayly 1993). Thus minimizing the expression

$$\sum_i (V_t(r_i) - V_q(r_i))^2 - \lambda(Q - \sum_A q_A) = \min \tag{34}$$

with respect to each q_a determines the `best' charges, subject to charge conservation. The first sum in the equation (34) runs over a set of grid points.

Calculation of analytical electrostatic potential is time consuming process. Rather than attempting to cover all space, it covers the valence regions of the molecule in which the electrostatic potential is expected to play an important role in chemistry. This region may be defined by a three-dimensional `envelope' around the molecule. The inner surface excludes points too close to the nuclei, and the outer surface excludes the regions which contain too little charge density to be significant. The inner surface of this envelope is typically defined by some van der Walls radii; the outer surface is a typically the van der Walls radii plus some value r. Within these two surfaces, grid points are generated on a regular rectangular mesh.

Since atoms can be represented by 3D spheres, a protein is usually modeled as a set of hard spheres, which is called a space-filling model, where radii are the van der Waals radii of atoms. A solvent accessible surface is the set of centers of a spherical probe rolling around the protein (Connoly 1983, 1985 Lee&Richards 1971). A probe is used for the computational convenience of a small molecule which interacts with the protein. A water molecule is frequently used as a solvent around a protein and the corresponding probe is approximated by a sphere with a radius of 1.4 Å .

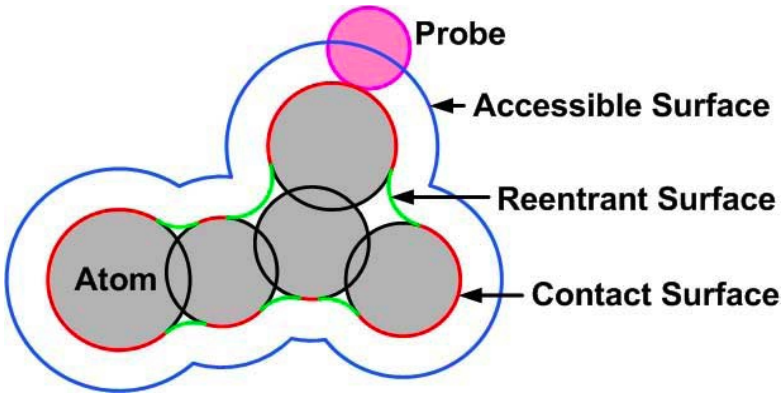


Figure 8 *`Envelope' around the molecule*. Solvent accessible surface and Connolly surface.

3.5.1 Application

Points were calculated at Connoly surface with density 0.5 points A^2. Points were calculated at three shells: 1) r_{vdw}+0.2Å, 2) r_{vdw}+0.4Å and 3) r_{vdw}+0.6Å, where r_{vdw} is van der Waals radii. Point surface density was 0.5/A^2. Electrostatic potential for molecule Thym was fitted in 856 point, Heps in 926 points and Crambin in 5114 grid points.

During fitting process some restraints were applied. The net charges of molecules and amino acids were zero, except -1.0 and +1.0 for first and last amino acid, respectively. Net charge for charged amino acids Arginin, Lysin was 1.0 and -1.0 for Glutatmat and Aspartat.

Standard force field parameters (partial point charges) were obtained with PHENIX-ELBOW (Adams 2002) package, and with pdb2pgr-server (Dolinsky 2004, 2007) for AMBER (Kollman *et al* 2008) parameters.

3.5.2 Measure of quality – linear correlation and relative root mean square

To compare electrostatic potentials we have used linear correlation coefficient (CC). This is statistical

procedure used to determine the degree to which two (or more) variables vary together. For the CC calculation all Connoly surface points were used. A CC greater than 0.8 is generally described as a strong correlation, whereas a CC less than 0.3 is a weak correlation. The correlation coefficient between two variables x_i and y_i where i = 1, 2, ..., n, is calculated as:

$$CC = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{35}$$

Linear correlation coefficients between various electron density maps and various electrostatic potentials were calculated. Global map CC is made between all points in unit cell which gives us information about quality of whole map. Linear correlation between various electrostatic potential was calculated using point as defined when fitting partial point charges.

The quality of the fit was calculated as relative root mean square (RRMS) given by:

$$RRMS = \left\{ \frac{\sum_i (V_t(r_i) - V_q(r_i))^2}{\sum V_t^2(r_i)} \right\}^{1/2} \tag{36}$$

Sum runs over a set of grid points (Connoly surface). Electrostatic potential calculated using electron density and protons is presented as V_t, and V_q is electrostatic potential calculated from partial point charges. Relative root mean square was used as target function for minimization. The geometry of the all molecules were based on the x-ray structural data.

3.5.3 Real space R-factor

Real space R-factor (Jones et al., 1991), between various Fc-type maps was calculated. Real space R-factor is a model-validation tool, and is calculated as follows:

$$RSR \frac{\sum (\rho_1 - \rho_2)}{\sum (\rho_1 + \rho_2)} \quad , \tag{37}$$

where ρ_1 and ρ_2 are maps which are compared. Sum runs over all real space grid points.

3.6 “Debugging”

To find out if our point charge procedure involve some systematically errors we decided to test fitting procedure for calculation of analytical electrostatic potential. The test-fitting procedure: First electrostatic potential using Mulliken point charges directly from GAMESS was calculated. This electrostatic potential was chosen as the target electrostatic potential. In next step all point charges were set to zero and point charge fitting procedure until convergence was applied. Used function for minimization was RRMS between potentials. Agreement between original charges and charges obtained with minimization where about ± 0.001, RRMS between potentials was bellow 1%. Second test was comparing electrostatic potential calculated using GAMESS and electrostatic potential calculated in MAIN. Theoretical quantum mechanical electron density was calculated using program GAMESS and introduced in MAIN. Then electrostatic potential was calculated using MAIN program. Potentials were compared and an almost perfect agreement was found, differences were below below 1%.

3.7 Poisson-Boltzmann equation

One of the most widely utilized approximate descriptions is the Poisson-Boltzmann equation, which can be motivated heuristically as a modification of Poisson's Equation, namely,

$$\vec{\nabla}(\epsilon(\vec{r})\vec{\nabla}\Phi(\vec{r})) = -4\pi[\rho(\vec{r}) + q\overline{n}_p e^{-\beta q \phi(\vec{r}) o(\vec{r})} - q\overline{n}_n e^{-\beta q \phi(\vec{r}) o(\vec{r})}] \tag{38}$$

The first term accounts for the fixed source charges (again, usually embedded in macroions). The other

two terms count, respectively, the contributions of the mobile positive and negative electrolyte ions. These are assumed to be in thermal equilibrium at temperature T ($\beta=1/kBT$). The potential energy of charge q located at position r is given by $q\Phi(r)$ where Φ is the electric potential. The quantities $n+$, $n-$ are the bulk number densities of cations and anions.

The Poisson-Boltzman equation is a differential equation that describes electrostatic interactions between molecules in ionic solutions. Electrostatics due to its long-range nature thought to be a key factor in characterizing biomolecular interactions. The methodology to determine electrostatic potential around molecules is often based on the Poisson-Boltzman mesoscopic theory describing the electrostatic potential and equilibrium distribution of ions around molecules in solution. In this model a molecule is represented as a set of partial charges immersed in implicit solvent with ions modeled by Boltzmann distribution. The resulting Poisson-Boltzmann equation is solved numerically on a three dimensional grid to obtain the electrostatic potential in each grid point.

3.7.1 Application

Poisson-Boltzman electrostatic potential was calculated with program CHARMM version „c36a1q“. All calculations were performed at Chemistry Institute Ljubljana. Below are shown some parameters which were introduced in to CHARMM input file:

```
set EpsR =   1  !!!  dielectric constant for the reference environment
set EpsP =   1  !!! dielectric constant for the protein interior
set  EpsW =   80  !!!  solvent dielectric constant
set Conc =   0.15 !!!  salt concentration
set Focus = 1  !!! to have a refined calculation focussed on the site using a finner grid
set Dcelc =  1.0  !!! the grid spacing in the finite-difference (centered on Xcen,Ycen,Zcen)
set Dcelf = 0.3 !!! the grid spacing in the finite-difference (centered on Xcen,Ycen,Zcen)
set LEdge  = 8   !!! distance between a protein atom and a grid
set Options =  watr 1.4 reentrant !!! Let's use the molecular surface
```

Input files for Crambin model were made with CHARMM-GUI (ref)input generator. Input files for Heps and Thym were made with the help of Milan Hodošček at Chemistry Institute Ljubljana.

3.8 Calculation of structure factors

Structure factors were calculated in three different ways:

- first (standard approach) using Anisotropic B-factor;
- second using wave function coefficients WFC;
- third (quantum mechanics) using electron density obtained from GAMESS

Structure factors were used for the F_{calc} (Fc) map and $2F_{obs}–F_{calc}$ (2FoFc) difference map calculation. Structure factors calculated in standard way were performed with SHELX (Shledrick 1990, 1993) for small molecules and PHENIX (Adams 2002) for protein model. Structure factors calculated using wave function coefficients were performed with MAIN (Turk 1992) for all models. Quantum mechanical electron density was calculated with GAMESS (Schmidt et al., 1993, Gordon and Schmidt, 2005) and Macmolplt (Bode&Gordon 1998) program.

3.9 Main chain charges

Amino acids are subunits of proteins. All of the 20 amino acids have in common a central carbon atom CA to which are attached a hydrogen atom, an amino group (NH2), and carboxyl group (COOH). These common atoms are also usually named as main chain atoms. The sum of all main chain charges (excluding hydrogen atoms) was defined as the main-chain charge.

4 Results

4.1 Kick maps – low resolution

The calculation of good quality electron density map of medium-to-high resolution is a crucial step in modern protein crystallography. We have improved maps quality by introduction of averaged kick maps.

4.1.1 Local map comparison along the chain

The quality of the AK maps was explored and compared with the final *Fmodel* map in parallel with the other available map types; UN, and ML and SA. The comparisons were performed on the local level, taken into account density points around selected regions of molecular models.

The comparison of maps locally at individual residues along the polypeptide chain revealed that ML and AK maps interchange in their ranking of similarity to the final *Fmodel* map. Since these maps were generated from the same structures from the middle of structure determination the fit of the residues along the chain fluctuates and so do the CC-s, which differ from map to map. The overlapping graphs of CC-s of UN, ML and AK maps along the whole chain are therefore not informative. To illustrate the differences between the AK and ML maps we have chosen two regions, in which the two kinds of maps are ranked differently. Figure 9b shows the region in which the ML map is closer to the final map, whereas the Figure 9a shows the region where the AK map is ranked highest. These graphs (Figure 9) reflect the match between the final model and electron density maps they are showing. The Figures (10a, 10b, 10c) correspond to the region of ammodytin L from the Figure 9a and the Figures (10d, 10e, 10f) correspond to the region of stefin B from the Figure 9b. The maps 10b and 10f have the highest match with the final F_{model} map. These comparisons illustrate that the use of AK maps in combination with ML maps can be useful during the model building procedures and are consistent with the past experiences.

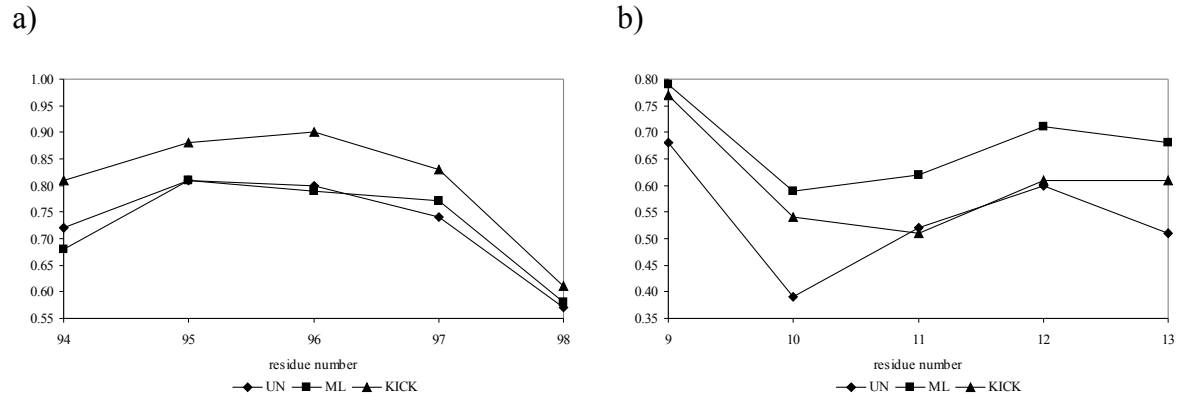


Figure 9 *Local map comparison in CC-s*. CC-s between the final Fmodel and initial UN, ML, AK maps using ammodytin L at 3.0 Å resolution. In (a) the results of Ile94-Glu98 residue region are shown while the (b) refers top the Ile9-Thr13 residue region.

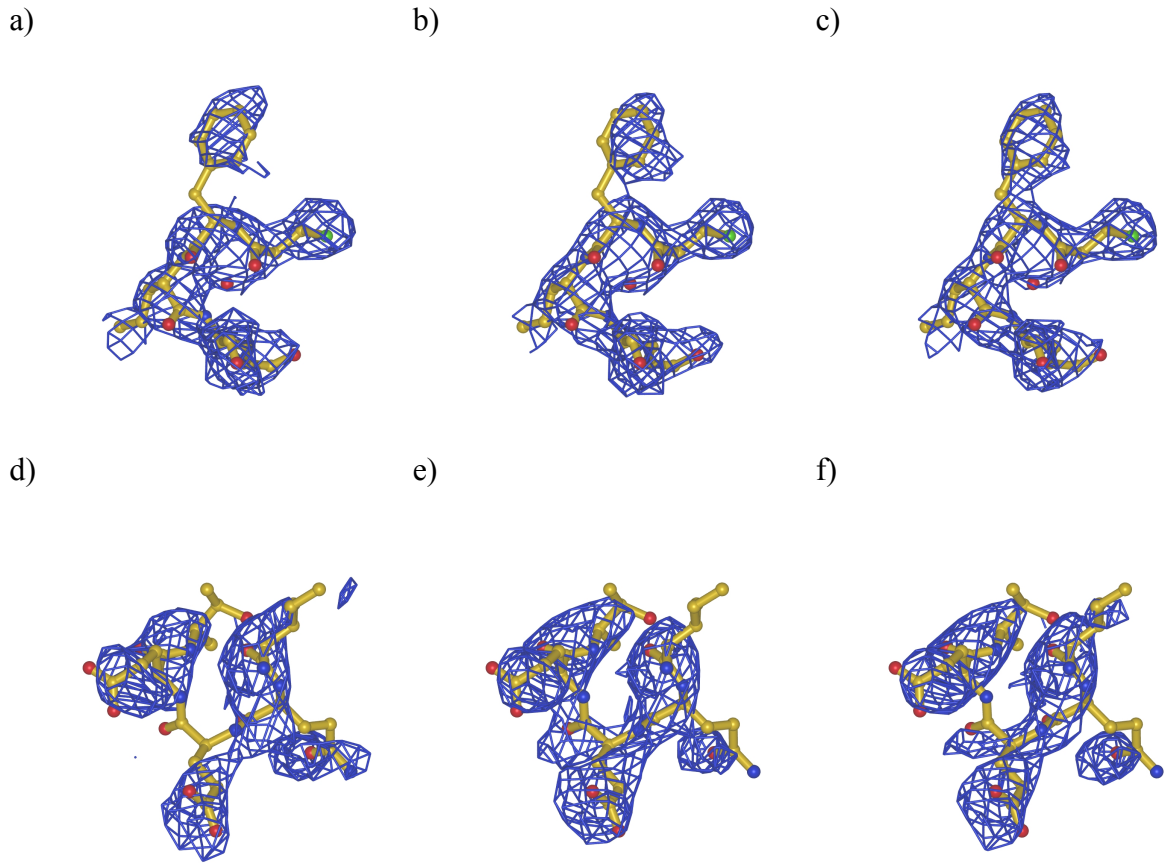


Figure 10 *Local map comparison in images*. The Figures (a, b, c) show maps around residues Cys94 – Arg98 while figures (d, e, f) show maps around residues Ile9 – Glu13 of the initial model of ammodytin L. The final model is shown in stick representation. Maps on figures (a, d) represent the UN maps, figures (b, e) the ML maps and figures (c, f) show the AK maps. The maps were generated using data at 3.0Å resolution and are all shown at 1.0σ contouring level. Kick map pics were made with MAIN and POVray program

4.2 Characteristic of Subatomic Resolution Refinement

At subatomic resolution, the ratio of the number of observations to the number of parameters (x, y, z coordinates, occupancy and wave function coefficients) is very high. Table 2 shows the number of observations (data) and the number of parameters when the models wave function coefficients using a TZV basic set. In all three cases, the ratio is high enough to use an unrestrained refinement.

Table 2 *Subatomic refinement*. For all models, the number of refined parameters as well as the amount of observable data and ratios are shown. The ratio is defined as observable/parameters. The ratio was calculated at final resolution and at 1.0Å resolution. The number of wave function coefficients (WFC) and the final resolution are also displayed.

Model	Resol(Å)	W.F.C	Parameters	Observable(final)	Ratio(final)	Observable(1Å)	Ratio(1Å)
Thym	0.38	280	404	11108	27	695	2
Heps	0.40	271	403	15839	39	1105	3
Cram	0.54	6665	9233	112233	12	18352	2

4.3 Small Molecules

At the very beginning when info on the values of wave function coefficients was unavailable, all S-type coefficients were set to 1.0, the only exception being hydrogen atoms where initial wave function coefficients were set to a value of 0.1. Orbitals of a higher order (P-type) were set to a zero value. These initial conditions give an R-factor of about 30%. Table 3 shows the refined values of wave function coefficient

Table 3 *Refined values of wave function coefficients*. Wave function coefficients for oxygen, carbon and two hydrogen atoms in the Thym molecule. Wave function coefficients were refined at a 0.38Å resolution, no restraints were applied. As shown in Table 3, in the first column are written consecutive numbers of wave function coefficients, in the second column are written atomic orbitals and in the third column are written wave function coefficient values. The atom name is written according to number-letter code, as in the PDB file.

Atom >>1	Orbital	O2
1	1S	0.49359
2	2S	0.45150
3	3S	1.05707
4	4S	0.95654
5	5S	0.49967
6	6Px	0.25216
7	6Py	0.01689
8	6Pz	-0.01755
9	7Px	0.26273
10	7Py	0.09660
11	7Pz	0.05833
12	8Px	0.20830
13	8Py	0.05329
14	8Pz	0.03305
Atom >>2		C8
85	1S	1.01078
86	2S	0.15520
87	3S	-0.04529
88	4S	-0.04987
89	5S	-0.04897
90	6Px	-0.06057
91	6Py	0.24677
92	6Pz	0.48611
93	7Px	0.02061
94	7Py	-0.00110
95	7Pz	0.09953
96	8Px	0.01400
97	8Py	-0.01657
98	8Pz	0.01587
Atom >>30		H13
275	1S	0.52022
276	2S	0.19450
277	3S	0.12898
Atom >>31		H14
278	1S	0.01266
279	2S	0.00113
280	3S	-0.00017

The success of an atomic model is often judged by the standard crystallographic R-factor, which is simply the average fractional error in the calculated amplitude compared to the observed amplitude.

Refinement at resolution 1.0A and lower, using wave function coefficients, yields a lower R-factor than when refined in a standard manner. In the case of the Thym R-factor, using wave function coefficients yielded a 1% lower value than when R-factor was calculated in the standard manner. At resolutions higher than 1A, the R-factor using wave function coefficients was as low as the R-factor calculated using the standard approach. As seen in Figure 11a, 11b refinement using wave function coefficients at a resolution higher than 0.6 Å yields a substantially higher R-factor than the standard approach. At final resolution, R-factor was 10.7% and 6.2% for Thym and Heps, respectively. For the sake of comparison, structure factors from theoretical quantum mechanical wave functions were calculated. The "quantum" R-factor was 39% and 27% for Thym and Heps, respectively. Very high "quantum" R-factors are due to the fact that theoretical models were applied to isolated (state of an ideal gas) molecules.

It was expected that refinement using wave function coefficients would yield a lower R-factor than when calculated using the standard approach. Namely, using TZV basic set, the electron density of "heavy" atoms is described with 14 contracted functions; on the other hand, the 6 parameters per atom are used for the anisotropic description. At this stage, the wave function coefficients method does not take into account thermal motion of atoms as in the standard approach. At a resolution below 1.0Å, the data parameter ratio is lower (the system is over-determined) and the temperature factor is compensating in wave function coefficients. At a resolution higher than 0.8Å, the ratio is much higher, therefore additional parameters are needed to take thermal motion of atoms into account.

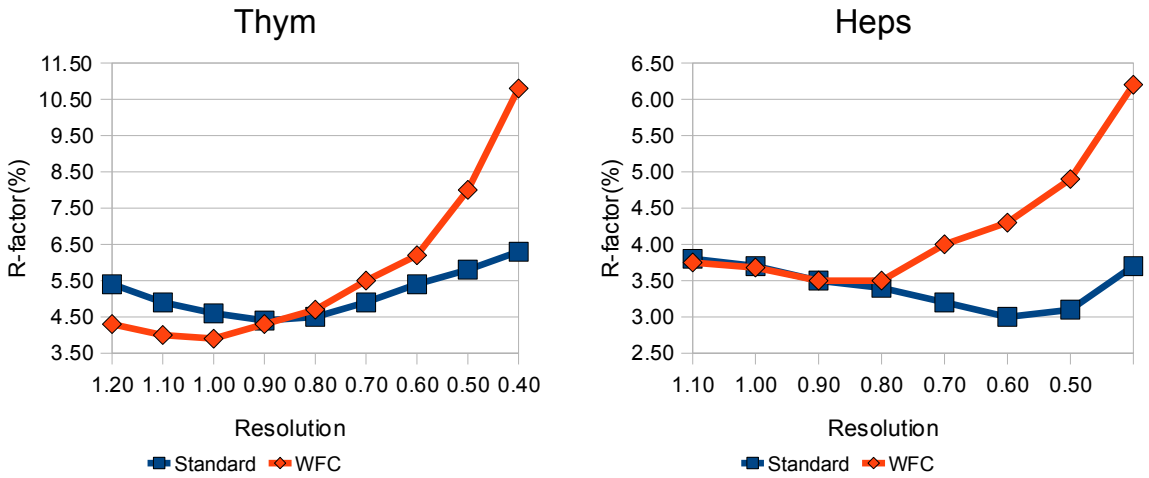


Figure 11 *Refinement at various resolutions.* R-factor value at different resolutions, calculated in the standard way and using refined wave function coefficients.

4.3.1 Comparison of electron density maps

To discover the similarity between various Fc-type maps, the linear correlation coefficient and real space R-factor were calculated. All maps are calculated at a maximum resolution of 0.38A and 0.50A for Thym (Figure 12) and Heps (Figure 13), respectively.

In the case of the Thym molecule, the highest correlation (CC=0.99) is between the Co_Fc and St_Fc maps, while the correlation of St_Fc with Co_Fc and Qm_Fc is approximately 20% lower. In the case of the Heps, the highest correlation (C=0.90) is shown by the pair Qm_map/Co_map, while the lowest correlation (CC=0.80) is between map Qm_map and St_map. In the first case (Thym), the St_Fc/Co_Fc pair show the largest similarity, in the second case (Heps), the Qm_Fc/Co_Fc pair show the highest correlation.

All maps show high linear correlation similarities, but on the other hand, high real space R-factor values show large differences between maps. It seems that the maps vary similar, however, they have very different mean values of electron density.

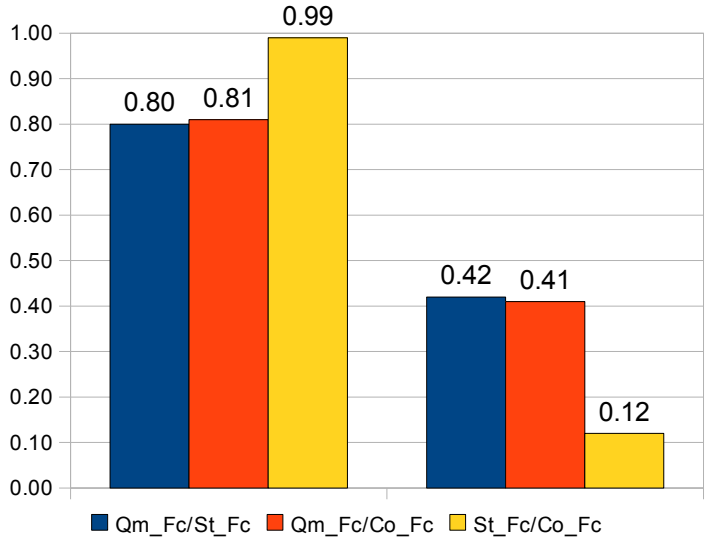


Figure 12 *Map comparison.* Linear correlation (left section) and real space R-factor (right section) between various Fc maps (Thym molecule) are displayed. Qm_Fc: structure factors were calculated from quantum mechanical electron density. St_Fc: structure factors were calculated using the standard approach Co_Fc: structure factors were calculated using wave function coefficients.

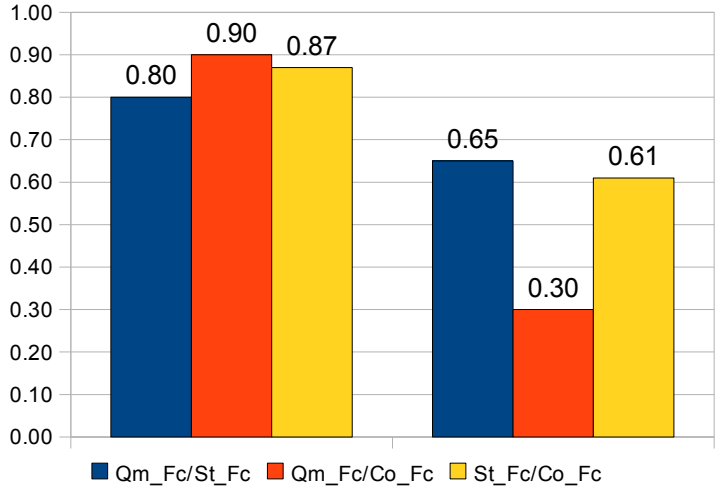


Figure 13 *Map comparison.* Linear correlation (left section) and real space R-factor (right section) for the Heps molecule are displayed. Values were calculated between various Fc type maps (see legend in Figure 12).

4.3.2 Point charge fitting

Point charges were fitted against various experimental and theoretical electrostatic potentials. All initial charges were set to a zero value, except in one case (marked with a star) where initial values were set to Mulliken charge values, calculated by the GAMESS program at HF/TZV level. As seen in Figure 14, all final energy RRMS were around 0.30 to 0.40. The lower value of RRMS occurred when fitted against Qm_Fc and St_2Fofc (RRMS=0.32). The highest value occurred at Co_2FoFc* potential. Similar conclusions can be made when looking at CC results. The highest correlation of 0.95 occurred during the case of Qm_Fc and St_2FoFc potential, the lowest CC was 0.91 at Co_2FoFc.

RRMS values in the case of the Heps molecule (see Figure 15) were slightly lower than in the Thym case. All RRMS span from 0.25 to 0.35. The lowest value of RRMS and the maximum value of CC occurred at Qm_Fc and Co_Fc. The highest value of RRMS occurred at St_2Fofc* (RRMS = 0.34, CC = 0.94). Results show that RRMS for both models span from 25% to 40%. Starting with zero initial values

yields lower (5%) RRMS than beginning with Mulliken charge values.
In most cases, fitting against Fc-type electrostatic potential gives lower RRMS, most likely because the Fc electron density maps involve less noise than 2FoFc.

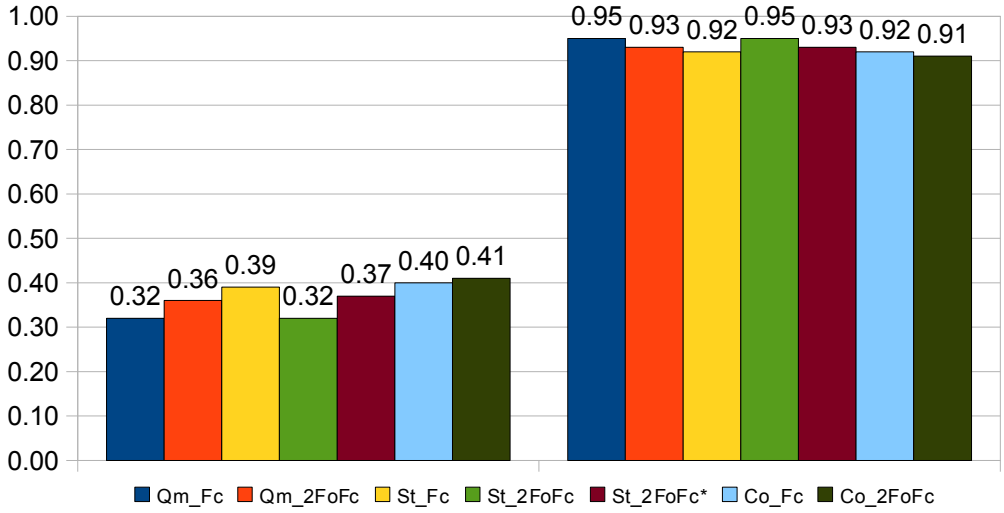


Figure 14 *Point charge fitting.* The results of point charge fitting for the Thym model. RRMS (left section) and CC (right section) between target electrostatic potential (calculated from electron density plus protons) and electrostatic potential calculated from fitted partial point charges are displayed. Qm_Fc and Qm_2FoFc: partial point charges were fitted against electrostatic potential using theoretical quantum mechanical structure factors and protons. St_Fc and St_2FoFc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_Fc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values.

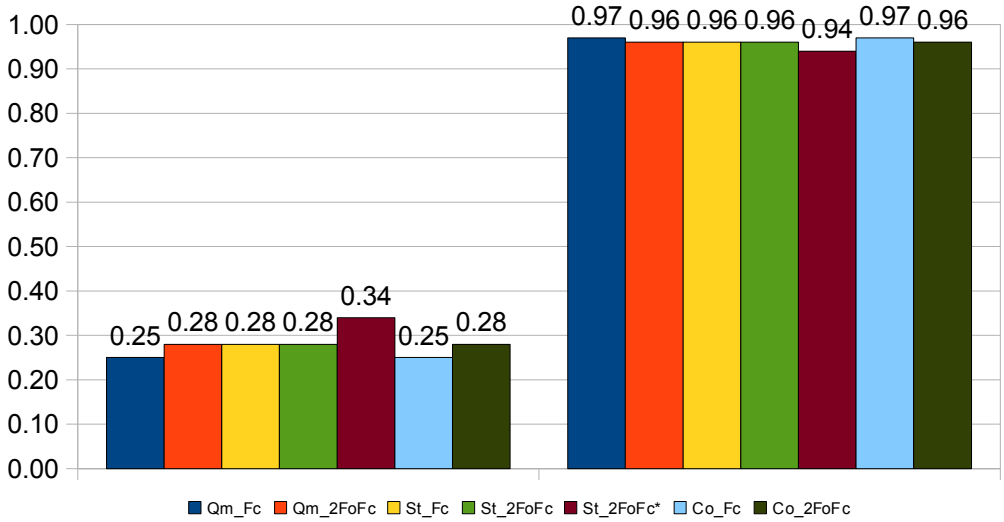


Figure 15 *Point charge fitting.* Results of point charge fitting for the Heps model. RRMS (left section) and CC (right section) between target electrostatic (potential calculated from electron density and protons) and electrostatic potential calculated from fitted partial point charges. For the legend, see Figure 14.

4.3.3 Linear correlation between various electrostatic potentials

Table 4 shows that linear correlation between GAMESS and ELBOW electrostatic potential is very high (CC = 0.94), but that they are in high anti-correlation with all other electrostatic potentials. Comparing electrostatic potentials which were fitted against experimental or theoretical electrostatic potentials yield high positive correlations. The most correlated (CC = 0.99) are pairs Qm_Fc/Qm_2Fofc,

St_2FoFc/St_2FoFc* and Qm_2FoFc/Co_2Fofc. It seems (Table 4) that the lowest correlation is between the potentials Co_Fc and Qm_Fc. In the case of the Heps molecule (Table 5), all potentials (except GAMESS and ELBOW) are in very high correlation (CC = 0.99 or 1.00).

The geometry of molecules was kept identical in all cases so that the resulting difference between the fitted electrostatic potentials must have originated in the distribution of electronic density.

Table 4 *Comparison of potentials.* For the Thym molecule, CC between various electrostatic potentials is displayed. All electrostatic potentials were calculated from a various set of point charges. Qm_Fc and Qm_2FoFc: partial point charges were fitted against electrostatic potential using theoretical quantum mechanical structure factors and protons. St_Fc and St_2FoFc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_Fc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. GAMESS: electrostatic potential calculated from Mulliken point charges at HF/TZV level. ELBOW: electrostatic potential calculated using ELBOW force field point charges.

CC	Qm_Fc	Qm_2FoFc	St_2FoFc	St_2FoFc*	GAMESS	St_Fc	ELBOW	Co_2FoFc
Qm_2FoFc	0.99	/						
St_2FoFc	0.96	0.98	/					
St_2FoFc*	0.96	0.98	0.99	/				
GAMESS	-0.56	-0.58	-0.66	-0.57	/			
St_Fc	0.93	0.94	0.94	0.97	0.43	/		
ELBOW	-0.47	-0.51	-0.62	-0.55	0.94	-0.41	/	
Co_2FoFc	0.97	0.99	0.98	0.98	0.60	0.93	0.55	/
Co_Fc	0.89	0.92	0.96	0.98	0.53	0.97	0.57	0.94

Table 5 *Comparison of potentials.* For the Heps molecule, CC between various electrostatic potentials is displayed. For the legend, see Table 4.

CC	Qm_Fc	Qm_2FoFc	St_2FoFc	St_2FoFc*	GAMESS	St_Fc	ELBOW	Co_2FoFc
Qm_2FoFc	1.00	/						
St_2FoFc	1.00	1.00	/					
St_2FoFc*	0.99	0.99	0.99	/				
GAMESS	-0.84	-0.84	-0.84	-0.79	/			
St_Fc	1.00	1.00	1.00	0.99	0.84	/		
ELBOW	-0.84	-0.83	-0.83	-0.78	1.00	-0.84	/	
Co_2FoFc	1.00	1.00	1.00	0.99	0.84	1.00	0.84	/
Co_Fc	1.00	1.00	1.00	0.99	0.84	1.00	0.84	1.00

4.3.4 Point charges

In order to understand the origin of the differences between electrostatic potentials, a comparison between partial point charges was made.

Tables 6 and 7 show the partial point charges calculated with ELBOW, GAMESS and point charges fitted against different experimental and theoretical electrostatic potentials. In the case of Thym (Table 6), charges of oxygen atoms do not differ significantly. The Thym molecule has two nitrogen atoms, however, the nitrogen charges calculated by ELBOW and GAMESS are negative (about -0.50), while in all other cases they are strongly positive (0.2 to 0.8). The difference in charges is also shown in carbon atoms. Carbon atoms C5, C7 are neutral when calculated with ELBOW and GAMESS, but C5 is strongly negative and C7 is positive when fitted against experimental and theoretical electrostatic potential. Atom C11 is neutral when fitted against electrostatic potential Qm_2FoFc, in all other cases it is strongly positive. Atom C14 is neutral when calculated with ELBOW, GAMESS or fitted against Qm_Fc potential, in all other cases it is strongly negative.

In the case of Heps (Table 7), molecule charge values of nitrogen, oxygen and hydrogen are quite similar. The Heps molecule also has one sulphur atom. The ELBOW point charge for S1 is +3.098, in all other cases it is around +2. Atoms C10 and C13 are neutral when calculated with ELBOW and GAMESS,

but the point charge value of C10 atom is strongly positive and the value of C12 is strongly negative when fitted against analytical electrostatic potential. As seen, atoms C6, C8, C11, C12, C14 have neutral charge values when calculated with ELBOW, whilst in all other cases point charge values are negative (-0.2 or less).

Table 6 *Point charges.* The table shows (Thym model) values of partial point charges calculated using different methods. Qm_Fc and Qm_2FoFc: partial point charges were fitted against electrostatic potential using theoretical quantum mechanical structure factors and protons. St_Fc and St_2FoFc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_Fc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. GAMESS: Mulliken point charges directly from GAMESS program at HF/TZV level. ELBOW: Point charges were calculated using ELBOW force field. The first column represents consecutive numbers referring to PDB file and atom type (H, C, O, N). In each row of the table, charges are colored. Charges with similar values are colored identically.

ATOM	ELBOW	GAMESS	Qm_Fc	Qm_2FoFc	St_Fc	St_2FoFc	Co_Fc	Co_2FoFc
1 O	-0.311	-0.575	-0.390	-0.507	-0.814	-0.377	-0.878	-0.485
2 O	-0.235	-0.419	-0.352	-0.374	-0.376	-0.327	-0.364	-0.362
3 N	-0.450	-0.645	0.239	0.403	0.533	0.264	0.807	0.704
4 N	-0.407	-0.429	0.811	0.714	0.811	0.787	0.834	0.737
5 C	0.039	0.072	-0.872	-0.720	-0.806	-0.867	-0.871	-0.859
6 C	0.271	0.291	0.797	0.718	0.822	0.783	0.843	0.750
7 C	0.026	0.026	0.820	0.781	0.832	0.802	0.818	0.769
8 C	-0.274	-0.094	-0.864	-0.307	-0.877	-0.798	-0.874	-0.853
9 C	-0.203	-0.619	-0.900	-0.900	-0.854	-0.867	-0.897	-0.890
10 C	0.136	0.027	0.768	0.368	0.798	0.793	0.845	0.715
11 C	0.436	0.598	0.809	0.042	0.77	0.517	0.814	0.697
12 O	-0.364	-0.500	-0.681	-0.401	-0.850	-0.372	-0.853	-0.898
13 C	0.492	0.696	0.524	0.707	0.649	0.288	0.828	0.715
14 C	0.004	-0.074	-0.074	-0.479	-0.859	-0.583	-0.758	-0.883
15 O	-0.289	-0.591	-0.851	-0.422	-0.548	-0.372	-0.641	-0.412
16 O	-0.335	-0.518	-0.567	-0.509	-0.713	-0.411	-0.779	-0.422
17 C	-0.082	-0.484	-0.895	-0.899	-0.868	-0.867	-0.884	-0.879
18 H	0.072	0.148	0.095	0.131	0.233	0.115	0.227	0.120
19 H	0.069	0.222	0.064	0.135	0.266	0.103	0.148	0.126
20 H	0.100	0.257	0.068	0.094	0.050	0.050	0.050	0.050
21 H	0.092	0.198	0.053	0.137	0.257	0.104	0.153	0.126
22 H	0.113	0.235	0.255	0.218	0.233	0.159	0.241	0.284
23 H	0.037	0.146	0.061	0.097	0.265	0.102	0.133	0.112
24 H	0.097	0.198	0.248	0.266	0.275	0.248	0.265	0.256
25 H	0.048	0.181	0.050	0.069	0.068	0.058	0.050	0.052
26 H	0.202	0.353	0.096	0.081	0.056	0.052	0.073	0.077
27 H	0.074	0.201	0.056	0.090	0.178	0.104	0.106	0.078
28 H	0.264	0.389	0.196	0.182	0.072	0.221	0.096	0.214
29 H	0.124	0.179	0.231	0.155	0.264	0.124	0.279	0.243
30 H	0.066	0.165	0.097	0.059	0.076	0.082	0.083	0.061
31 H	0.187	0.367	0.109	0.071	0.059	0.084	0.105	0.055

Table 7 *Point charges.* The table shows (Heps model) values of partial point charges calculated using different methods. For details, see the legend in Table 6. The first column represents consecutive numbers referring to PDB file and atom type (H, C, O, N). In each row of the table, charges are colored. Charges with similar values are colored identically.

ATOM	GAMESS	ELBOW	Qm_Fc	Qm_2FoFc	St_Fc	St_2FoFc	Co_Fc	Co_2FoFc
1 S	1.641	3.098	1.991	1.995	2.000	1.992	1.994	1.992
2 N	-0.247	-0.332	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
3 N	-0.534	-0.126	-0.661	-0.631	-0.460	-0.668	-0.662	-0.666
4 O	-0.790	-1.125	-0.315	-0.314	-0.314	-0.315	-0.313	-0.314
5 O	-0.595	-0.334	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
6 C	-0.243	-0.051	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
7 C	-0.620	-0.932	0.868	0.870	0.872	0.868	0.869	0.868
8 C	-0.183	-0.041	-0.471	-0.288	-0.576	-0.532	-0.530	-0.547
9 O	-0.795	-1.118	-0.315	-0.314	-0.314	-0.315	-0.313	-0.314
10 C	-0.096	0.093	0.450	0.299	0.245	0.581	0.506	0.576
11 C	-0.166	-0.046	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
12 C	-0.206	-0.041	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
13 C	-0.093	0.048	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
14 C	-0.240	-0.006	-0.625	-0.631	-0.617	-0.631	-0.626	-0.629
15 O	-0.773	-1.092	-0.315	-0.314	-0.314	-0.315	-0.313	-0.314
16 H	0.148	0.074	0.145	0.142	0.147	0.142	0.144	0.143
17 H	0.219	0.117	0.145	0.142	0.147	0.142	0.144	0.143
18 H	0.186	0.053	0.145	0.142	0.147	0.142	0.144	0.143
19 H	0.220	0.132	0.145	0.142	0.147	0.142	0.144	0.143
20 H	0.215	0.110	0.145	0.142	0.147	0.142	0.144	0.143
21 H	0.168	0.068	0.145	0.142	0.147	0.142	0.144	0.143
22 H	0.194	0.105	0.145	0.142	0.147	0.142	0.144	0.143
23 H	0.366	0.212	0.145	0.142	0.147	0.142	0.144	0.143
24 H	0.215	0.130	0.145	0.142	0.147	0.142	0.144	0.143
25 H	0.395	0.246	0.145	0.142	0.147	0.142	0.144	0.143
26 H	0.225	0.113	0.280	0.281	0.281	0.281	0.282	0.281
27 H	0.255	0.103	0.145	0.142	0.147	0.142	0.144	0.143
28 H	0.203	0.122	0.280	0.281	0.281	0.281	0.282	0.281
29 H	0.219	0.113	0.280	0.281	0.281	0.281	0.282	0.281
30 H	0.228	0.126	0.280	0.281	0.281	0.281	0.282	0.281
31 H	0.160	0.063	0.145	0.142	0.147	0.142	0.144	0.143
32 H	0.153	0.078	0.145	0.142	0.147	0.142	0.144	0.143
33 H	0.173	0.038	0.145	0.142	0.147	0.142	0.144	0.143

4.3.5 Visual inspection – point charge electrostatic potential

Electrostatic potential generated by Thym is shown in Figure 16. The Co_2Fofc and Co_2FoFc(target) potentials compare well in the positive potential, but the negative region of Co_2FoFc(target) appears significantly stronger than the Co_2FoFc. The agreement between the Co_Fc and St_Fc potentials is generally good: the positive and negative potential regions are similar, but the St_Fc positive potential wells are generally deeper. AMBER charges yield similar potential to analytical quantum mechanical electrostatic potential (GAMESS-Macmolplt). As seen in Figure 16, strong anti-correlation exists between electrostatic potentials calculated using semi-empirical or theoretical methods and potential calculated using point charges fitted against analytical experimental electrostatic potential.

Electrostatic potential generated by Heps is shown in Figure 17. The Co_2Fofc and Co_2FoFc(target) potentials compare well in the positive potential, but the negative region of Co_2FoFc(target) appears less

strong than the Co_2FoFc. Figure 17 shows almost perfect agreement between Co_Fc and St_Fc potentials. Semi-empirical and theoretical potentials are in anti-correlation with potentials calculated from point charges fitted against analytical experimental electrostatic potential.

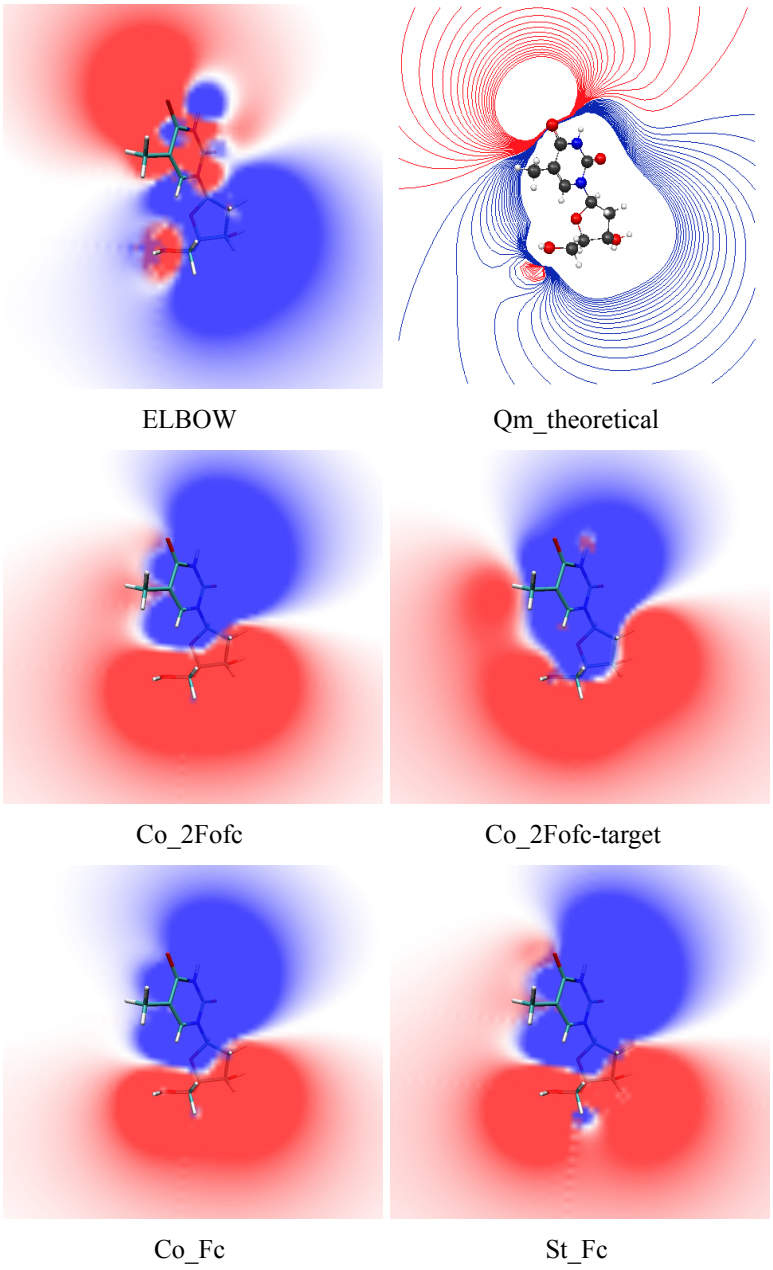


Figure 16 *Electrostatic potential generated by Thym.* All contour levels are 20 kT/e (T=300K). Blue -positive; red - negative. The electrostatic potentials of St_Fc, Co_Fc and Co_2FoFc were calculated from fitted point charges. St_Fc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_2FoFc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. Co_2Fofc-target: electrostatic potential calculated using wave function coefficient structure factors and protons. ELBOW: electrostatic potential calculated from ELBOW force field point charges. Qm_theoretical: theoretical electrostatic potential calculated using GAMESS at TZV/HF level and Macmolplt program. All electrostatic potential pictures, except Qm_analytical were made using VMD-openGL (Humphrey 1996) software. The Qm_analytical potential picture was made using Macmolplt (Bode & Gordon 1998).

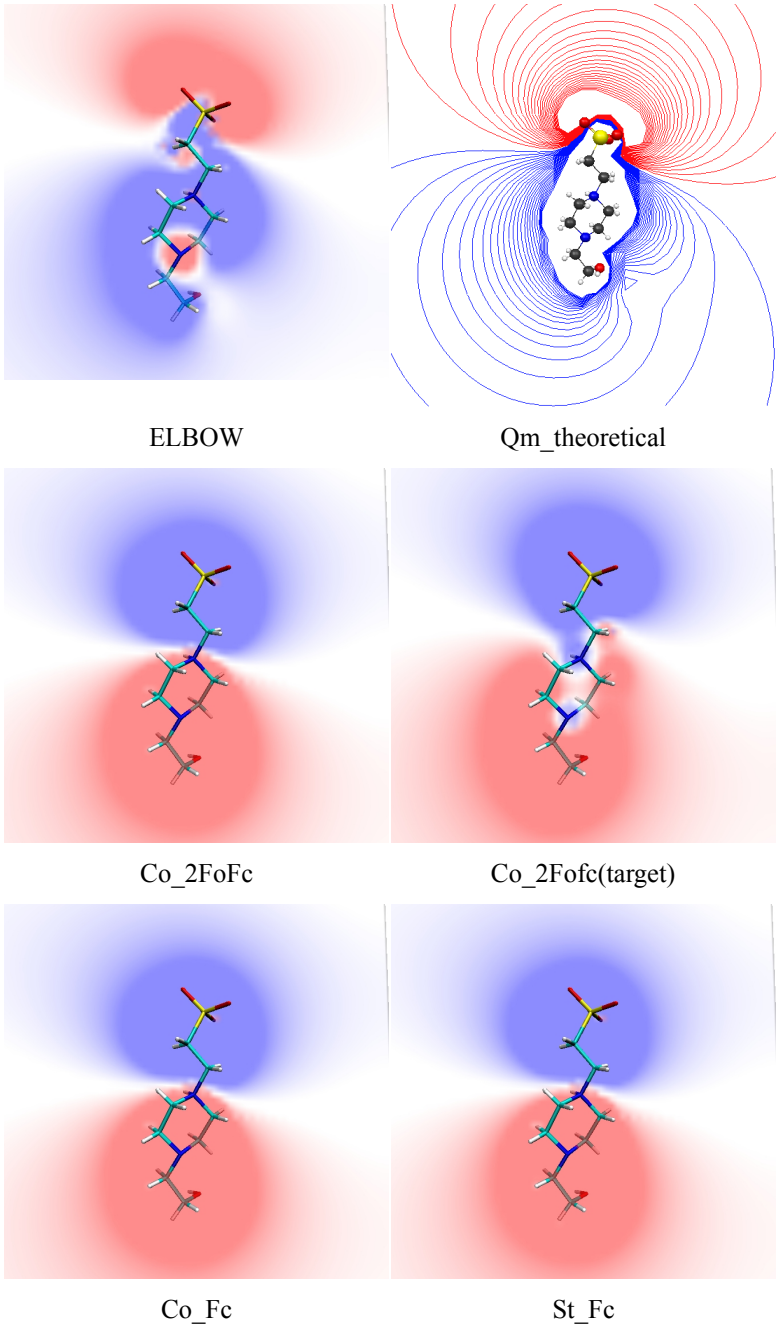


Figure 17 *Electrostatic potential generated by the Heps*. All contour levels are 20 kT/e (T=300K). Blue - positive; red – negative. For the legend, see Figure 6. All electrostatic potential pictures, except Qm_analytical were made using VMD-openGL software. The Qm_analytical potential picture was made using Macmolplt.

4.4 Protein model – Crambin

4.4.1 Point charge fitting

The procedure for point charge fitting was applied to a much larger model, namely the protein crambin (Figure 18). The initial charges were set to a zero value when point charges were fitted against St_Fc and St_2Fofc electrostatic. In the case where point charges were fitted against St_Fc electrostatic potential, the RRMS value was 0.35. In the second case, partial point charges were fitted against St_2FoFc electrostatic potential and the final optimized RRMS was 1.56 (=156%). This result suggests that the optimization has not converged. In the next step, the optimization procedure was repeated using Mulliken charges as initial values. Mulliken charges were calculated with the program GAMESS at HF/3-21G level. Calculated with

Mulliken initial values, the final RRMS dropped to 0.35 value. In addition, Mulliken charges were also set as initial values when fitted against St_Fc electrostatic potential. In this case, RRMS was 10% lower (RRMS=0.25) than in the case of zero initial values. Similar conclusions emerge when comparing the CC-s - the highest correlation was in the case of St_Fc* (0.97) and the lowest in the case of St_2FoFc (CC=0.40). These results show that point charge fitting optimization is dependent on initial values. Initial values play an important role when fitting point charges against experimental or theoretical electrostatic potential. Electrostatic potential St_2FoFc that started with zero charges did not converge, meanwhile the St_2FoFc* quality-of-fit was 35%.

The results of Figure 18 lead us to the next question, namely why minimization against St_Fc electrostatic potential is so much better then against St_2FoFc. In both cases, initial values of point charges were set to zero. One explanation could be that the 2FoFc-type map is nosier and minimization did not converge. The second assumption regarding the large differences is that crambin is much larger than smaller test models (Thym and Heps) and is also therefore a much more complex system. When using Mulliken charges as initial values, RRMS dropped dramatically. We can conclude that when fitting point charges for large systems such as proteins, initial values play a very important role. It should be noted that introducing Mulliken charges as initial values for small molecule models (Thym, Heps) results in slightly higher RRMS than with zero initial values.

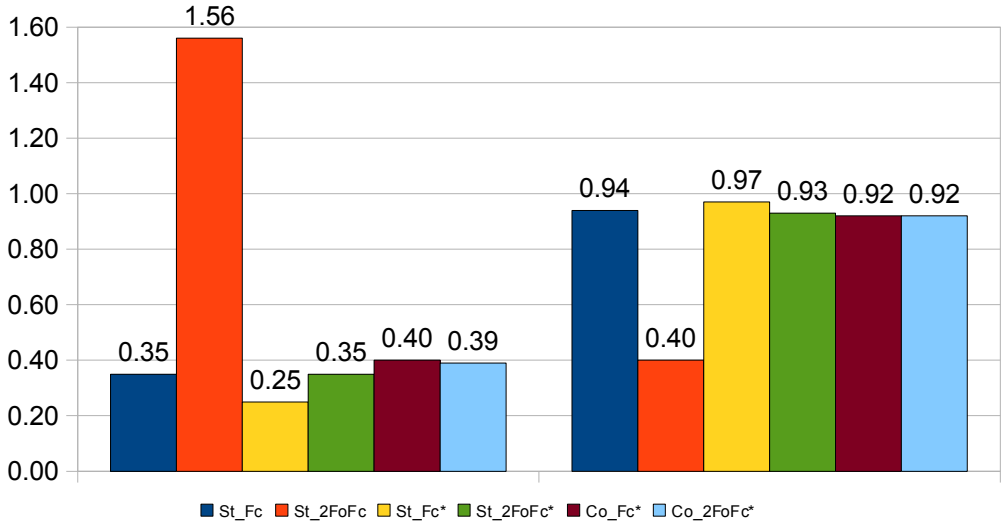


Figure 18 *Point charge fitting.* Results of point charge fitting for the crambin model. RRMS (left section) and CC (right section) between target electrostatic potential (calculated from electron density plus protons) and electrostatic potential calculated from fitted partial point charges is shown. Qm_Fc and Qm_2FoFc: partial point charges were fitted against electrostatic potential using theoretical quantum mechanical structure factors and protons. St_Fc and St_2FoFc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_Fc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values.

4.4.2 Linear correlation between various electrostatic potentials

In this section, the correlation between various electrostatic potentials is presented. All electrostatic potentials were calculated from partial point charges. Table 8 shows that St_2FoFc (RRMS=1.56, see Figure 18) has the highest correlation with AMBER and GAMESS electrostatic potential; with all other potentials, St_2FoFc is in low anti-correlation. When point charges were fitted against St_Fc and St_Fc* potential initial conditions did not play a significant role, since the two St_Fc and St_Fc* are in very high correlation (CC = 0.98). As seen in Table 8, initial conditions played an important role when charges were fitted against 2FoFc-type maps, namely potentials St_2FoFc and St_2FoFc* are in medium anti-correlation.

Table 8 *Comparison of potentials.* The linear correlation between various electrostatic potentials is shown. All electrostatic potentials were calculated from a various set of point charges. St_Fc and St_2FoFc: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_Fc and Co_2FoFc: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. GAMESS: electrostatic potential calculated from Mulliken point charges at HF/3-21G level. AMBER: electrostatic potential calculated using AMBER force field point charges.

CC	St_2FoFc	St_Fc	St_Fc*	GAMESS	St_2FoFc*	AMBER	Co_2FoFc*
St_Fc	-0.15	/					
St_Fc*	-0.14	0.98	/				
GAMESS	0.85	-0.05	0.03	/			
St_2FoFc*	-0.32	-0.09	-0.08	-0.43	/		
AMBER	0.84	0.07	0.03	0.94	0.30	/	
Co_2FoFc*	-0.32	-0.30	0.31	-0.43	0.89	-0.30	/
Co_Fc*	-0.15	0.86	0.89	-0.04	-0.33	0.02	0.27

4.4.3 Main-chain charge

Figure 19 shows that AMBER main-chain charges are roughly constant and vary depending on the type of amino acid. The first amino acid THR1 has a main-chain charge of +0.23, the last amino acid ASP46 has a main-chain charge of -0.6. All other amino acids have main-chain charges between -0.27 and -0.51. The main-chain charges of St_2FoFc* and Co_2FoFc* dramatically fluctuate between +2.0 and -2.5 values. The lowest values of main-chain charges were about -2.0 (ARG10, ILE35, CYS40) and the highest values were about +2.0 (ILE7, LEU18). A closer observation of Figure 19 shows that the main difference (more than +-1.0) between the main-chain charge of St_2FoFc* and Co_2FoFc* are in amino acids CYS3, CYS4, PHE13, ASN14, PRO19, TYR29.

In the presented model, point charges on atomic sites are allowed to fluctuate. Force fields, such as CHARMM or AMBER, which are used for the refinement of protein structures, have constant partial atomic charges irrespective of the chemical environment, such as secondary structure, solvent ... In reality, molecular electronic structures can be strongly influenced by the external environment.

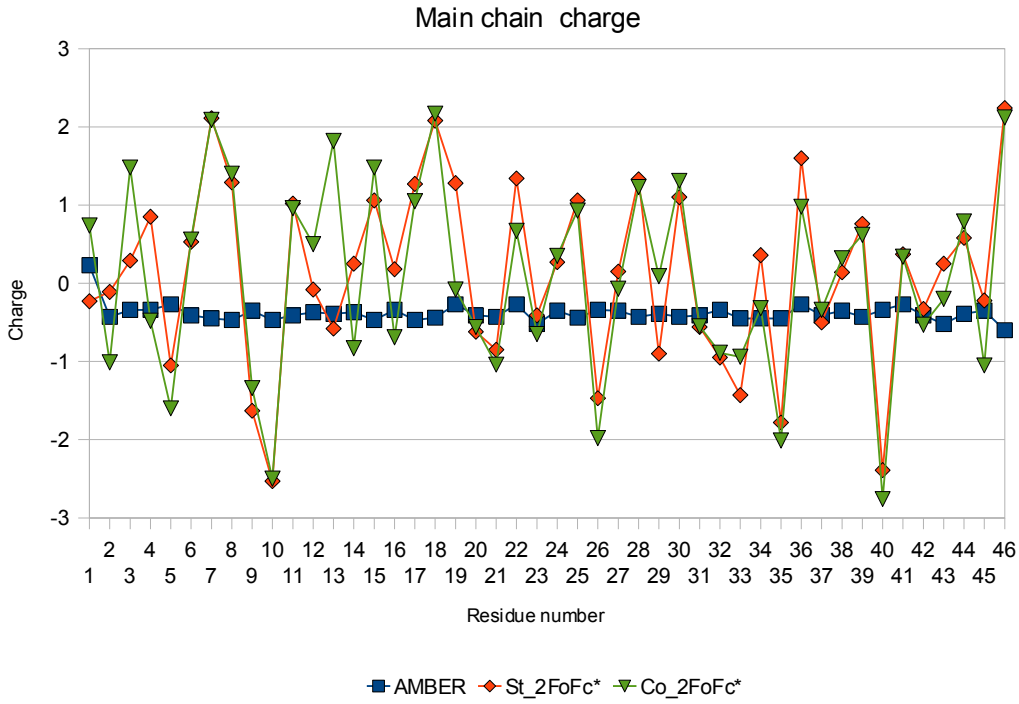
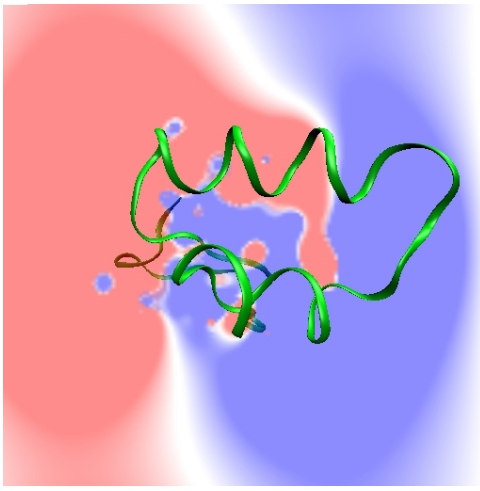


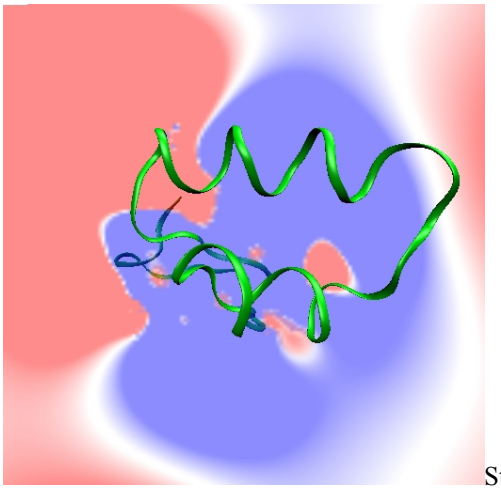
Figure 19 *Graph of main-chain charge values.* Main-chain charge is the sum of four main chain atoms; N, CA, C and O. Main-chain charge were calculated using different methods; AMBER force field parameters, point charges fitted against St_2Fofc* and Co_2FoFc* electrostatic potentials. (*) - fitting was started using Mulliken charges as initial values.

4.4.4 Poisson-Boltzmann electrostatic potential

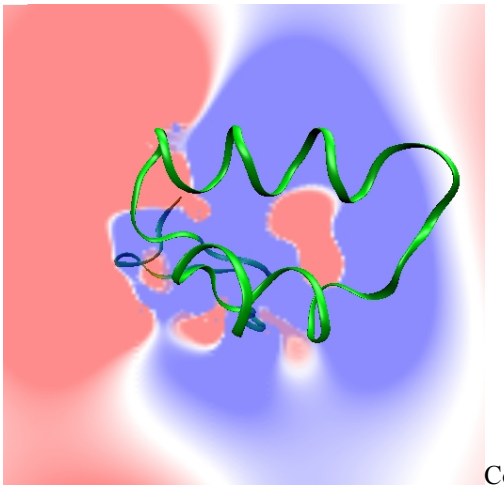
Electrostatic potentials with different sets of partial point charges using the Poisson-Boltzmann method is presented. Despite large differences between various main chain charges (Figure 19), Poisson-Boltzman electrostatic potentials do not differ significantly. In x-plane AMBER derived electrostatic potential is more positive around loop which connects alpha helixes and expose more negative around the longer helix. In plane-y, AMBER derived electrostatic potential is more negative then Co_2FoFc* an St_2FoFc*. Figure 20 shows (z-plane) that in vicinity of both alpha helixes AMBER electrostatic potential is less negative as Co_2FoFc* and St_2FoFc* potentials. The AMBER electrostatic potential is comparable in sign with Co_2FoFc* and St_2FoFc*, however is significantly different in magnitude. Much stronger positive and negative regions of Co_2FoFc* and St_2FoFc* potential comparing to AMBER is consequence of large charge fluctuations. Namely, fitted point charges fluctuate dramatically comparing to AMBER derived charges (see Figure 19).



AMBER (x-plane)

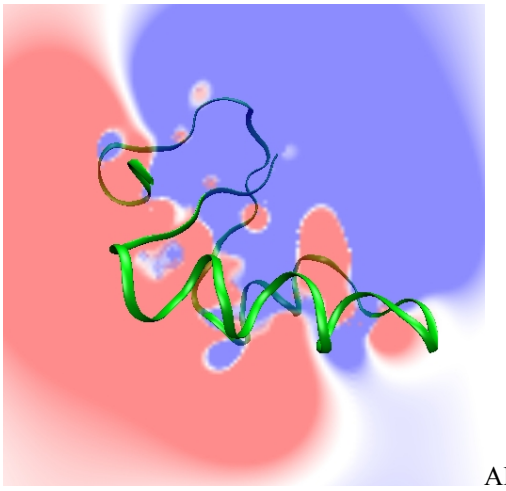


St_2FoFc* (x-plane)

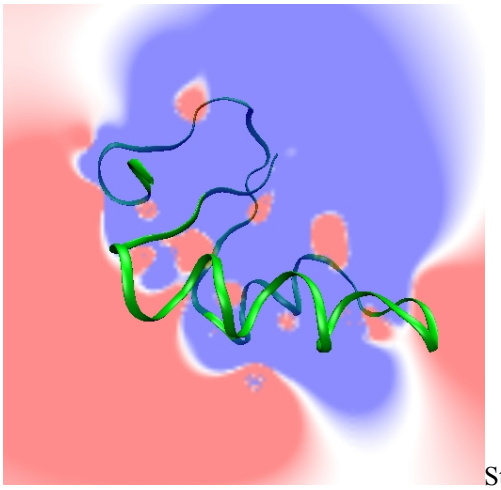


Co_2FoFc* (x-plane)

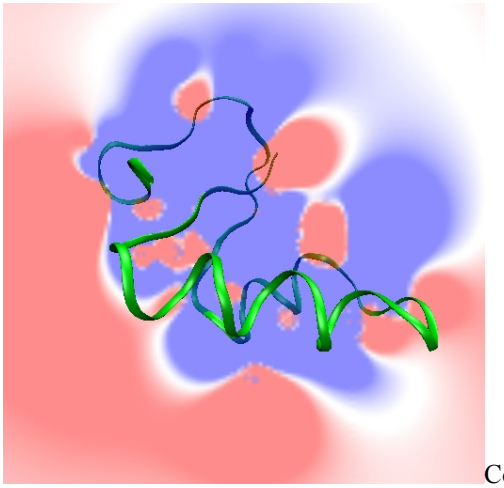
Figure 20 *Electrostatic potential generated by crambin in x plane (offset = 0.50) is displayed.* The Poisson-Boltzmann approach was used with different sets of partial charges. St_2FoFc*: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_2FoFc*: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. AMBER: electrostatic potential calculated using AMBER force field point charges. Contour level for AMBER and Co_2FoFc*, Qm_2FoFc* are 20kT/e and 60kT/e (T=300K), respectively. Blue - positive; red - negative. All electrostatic potential pictures were made using VMD-openGL software.



AMBER (y-plane)

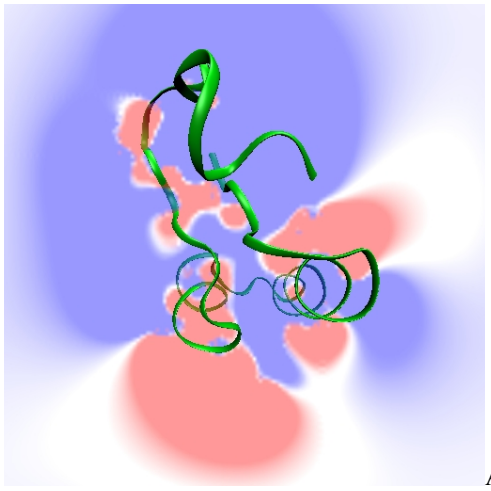


St_2FoFc* (y-plane)

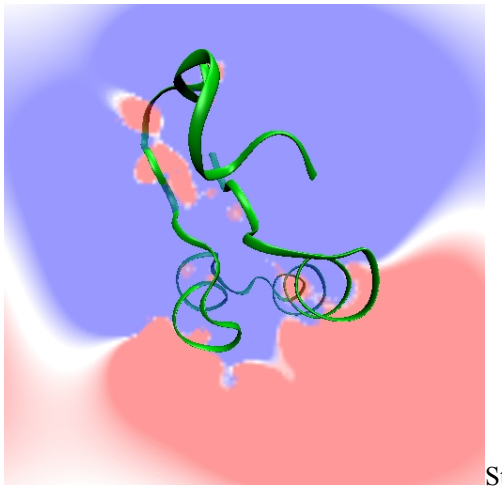


Co_2FoFc* (y-plane)

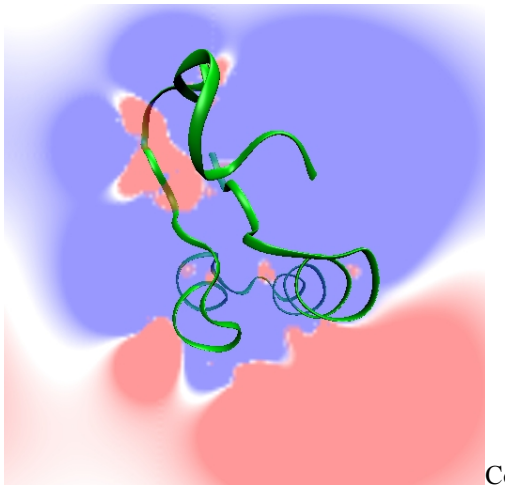
Figure 21 *Electrostatic potential generated by crambin in y plane (offset = 0.50) is displayed.* The Poisson-Boltzmann approach was used with different sets of partial charges. St_2FoFc*: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_2FoFc*: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. AMBER: electrostatic potential calculated using AMBER force field point charges. Contour level for AMBER and Co_2FoFc*, Qm_2FoFc* are 20kT/e and 60kT/e (T=300K), respectively. Blue - positive; red - negative. All electrostatic potential pictures were made using VMD-openGL software.



AMBER (z-plane)



St_2FoFc* (z-plane)



Co_2FoFc* (z-plane)

Figure 22 *Electrostatic potential generated by crambin in z plane (offset = 0.50) is displayed.* The Poisson-Boltzmann approach was used with different sets of partial charges. St_2FoFc*: partial point charges were fitted against electrostatic potential using standard structure factors and protons. Co_2FoFc*: partial point charges were fitted against electrostatic potential using wave function coefficient structure factors and protons. (*) – the fitting started by using Mulliken charges as initial values. AMBER: electrostatic potential calculated using AMBER force field point charges. Contour level for AMBER and Co_2FoFc*, Qm_2FoFc* are 20kT/e and 60kT/e (T=300K), respectively. Blue - positive; red - negative. All electrostatic potential pictures were made using VMD-openGL software.

5 Discussion

3-dimensional structure (macro)molecules are the interpretation of the experimental data. When crystallized molecules are available, the interpretation of the crystal diffraction pattern is obtained from an X-ray scattering (more rarely, particles such as neutrons and electrons are used) on the atoms in the crystal. The structure is described by a mathematical model, which more or less accurately describes the molecular structure. Hence, the mathematical model provides a more or less accurate interpretation of the measured data. Accuracy is defined as the ratio of the measured data and the parameters of the mathematical model which define the structure. The accuracy and the choice of the model are therefore defined as the amount of measured data which is directly reflected in the resolution of the diffraction pattern. The lower the resolution, the smaller the amount of experimental data (observed structure factors, "Fobs"); thus, a simpler model should be used for interpretation and vice versa.

The standard mathematical model of structure factors, which describes the electron density of molecules in the crystal, is derived from the spherical Gaussian functions. Coefficients of spherical Gaussian functions have been fitted to the electron density of atoms calculated from the Hartree-Fock wave function (Cromer & Mann, 1967). The 9 Cromer-Mann coefficients were calculated for numerous atoms in ground and ionic states. The Cromer-Mann coefficients are constant regardless of the structure of molecules in which individual atoms are located. Gaussian spherical functions are centered in the nuclei of atoms and thus determine the position of each atom in the molecule with positional parameters (x, y, z). The geometry (3D structure) of the molecule is described as the sum of Gaussian functions. At a resolution below 3Å positional, (x, y, z) parameters are actually the only adjustable parameters of molecular structures. Furthermore at a resolution of 3Å, the number of measured parameters is approximately three times the number of atoms. The accuracy of the position of each atom at resolution 3Å is about 1Å. By increasing the resolution, the amount of data increases. An individual isotropic factor (B-factor) is added to the three spatial coordinates (Konnert 1980, Willis 1975). The temperature factors or B-values, as determined by crystallography, are linearly related to the mean square displacement of an atom and give an indication of atom flexibility in the crystalline state. The mathematical model of the B-factor is also a Gaussian function (equations 7 and 8). The four parameters (x, y, z and B) well describe models up to a 2Å resolution (about 7 measured parameters at each atom). At a higher resolution, due to the large number of observables (typically on the order of 30 to 50 reflections per non-hydrogen atom), the isotropic model can be upgraded to an anisotropic model - 6 parameters to describe the orientation and the elongation of an ellipsoid (Sheldrick 1990, 1996, Dauter 1995, Dunitz, 1998). In this case, the Debye–Waller factor can be expressed as a symmetric 3×3 matrix B_j (equation 9). Anisotropic B-factors may be gradually added to individual atoms - first to heavier atoms such as sulphur and then to carbon, nitrogen and oxygen atoms. Spherical scattering factors well describe atom models of macromolecules to the resolution of approximately 1.2Å. At higher resolution levels (0.5Å - about 300 measured parameters at each atom), it is possible to obtain residual maps which clearly show valence electrons, including those of hydrogen atoms. The electron cloud around the atoms of a molecule is deformed as a result of chemical bonding and non-bonding interactions (notably hydrogen bonds) between the atoms.

The presence of significant densities in deformation maps at high resolution is an indicator that spherical free atom models, even with anisotropic thermal parameters, give too crude an approximation to the real electron density. The multipolar model (Hansen and Coppens 1978, Coppens 1997, Spackman 1998) which is widely used in small-molecule crystallography, gives a better representation of electron charge densities. The Slater functions are used to describe the spherical distribution of electron density, however the non-spherical electron density is described using spherical harmonic functions. Such a multipole analysis is necessarily limited by its very construction in the types of predictions it can make. The multipolar model is limited by the description of the distribution of electron density (as the standard model). The diffraction data at ultra high resolutions contains information about electron clouds, occupation of orbitals and electron shifts. This information allows comparability of experimental data with the theoretical quantum mechanical models. Therefore, it makes sense to use mathematical models for the determination of the structures at ultra high resolutions which are comparable (or similar) to quantum mechanical models. The standard electron density of a molecule is calculated as the sum of individual atom densities, using the Gaussian spherical scattering factor model. These Gaussian functions are

approximations of quantum mechanical wave function. Conceptual differences between standard scattering factors and wave function exist. The theory of quantum mechanics says that an electron's position can only be described statistically. It also describes a method for calculating the probability of finding an electron at one point or another. This calculation produces a quantity called electron density. The wave function is also called a probability amplitude because it is the square of the wave function that yields probabilities. In practice, the quantum mechanics equations have only been accurately solved for one electron system. The electronic wave function for many electron systems is a Slater determinant based on occupied molecular orbitals which are approximated by a linear combination of atomic orbitals.

5.1 Electron density

An accurate description of electron density is crucial for the computation of molecular properties such as experimental electrostatic potential. In order to overcome limitations of standard models, we have used the basis functions as used in the quantum mechanical models. We have named these functions “wave function coefficients” (WFC). For a description of electron density, we have used basis functions that satisfactorily describe the S, P and D orbitals. This approach also allows for a more accurate “picture” of the electronic density of small molecules as well as macromolecules. WFC electronic density is calculated from one “quasi-molecular orbital”, which is the sum of all atom orbitals in the molecule. In the case of the WFC model, electron density cannot be divided to individual atom electron densities. Therefore, the value in each grid (real space) was calculated and following this, each grid value was squared. X-rays are scattered at electron density, which in the quantum mechanical model is calculated as the sum of squares of all occupied molecular orbitals. Each molecular orbital is calculated as the sum of all atom orbitals. Referring to the WFC model, X-rays are scattered at the “quasi-molecular orbital” of the molecule. The standard model is based on the assumption that X-rays are scattered at individual atom densities. Individual terms in the squared “quasi-molecular orbital” do not represent individual atom densities. The WFC model allows the use of FFT algorithms, therefore the computation of mixed terms was not necessary. Concomitant use of analytical FT and WFC methods would require the development of a product of all terms for calculation of structure factors. In the case of the crambin protein model, approximately 6,600 atomic orbitals were used to describe electron density. The square of such a function will give 6,600*6,600/2 terms, approximately 22 million terms. It should be noted that crambin is one of the smallest proteins. It only has 46 amino acids, so the use of WFC and analytical FT at the same time for large molecules, such as those of proteins, is not suitable.

5.2 WFC model validation

A comparison of R-factors calculated using the standard approach and the WFC method shows that at a resolution of 1.0Å and lower, the WFC method gives a lower R-factor than the standard approach. At resolutions higher than 1Å, R-factor using standard structure factors gives lower values than the WFC method. The classic model better matches the measured data at higher resolutions, we assume that refinement using the WFC model at a lower resolution was over-determined. The ratio between measured data and refined parameters is low (about 2, for details see Table 2). It seems that the WFC model compensates thermal movements of atom nuclei. At high resolutions (above 0.8A), the mathematical model (WFC) of atomic orbitals is no longer capable of fully compensating thermal movements of atomic nuclei. This can be observed as a higher R-factor. The problem is of course, also the opposite; it is clear that the classic model cannot sufficiently describe electrons in valence orbitals. The standard model is also unappropriated for describing electron shifts in the molecule, however it seems that the standard model is able to partially compensate electron shifts. In order to more accurately describe the electron density using the WFC method at ultra-high resolutions, it is necessary to include features that will be able to adequately compensate the distribution of thermal movements of atom nuclei. Standard structure factors involve B-factors which represent smearing of atom electron densities around their equilibrium positions due to thermal motion and positional disorder. Because of conceptual differences, direct use of standard individual temperature factors in the formulation of WFC is not appropriate. One possible solution is to use a set of “kicked” models, as were used in the calculations of the electronic density map (Pražnikar et al, 2009). Instead of the use of the single position of an atom, multiple positions for a single atom can be used while using only one set of coefficients.

5.3 Electrostatic interactions

Electrostatic interactions play a crucial role in determining the structure and behavior of proteins and more complex structures, e.g. enzymes and viruses (Honig and Nicholls, 1995; Davis and McCammon, 1990; Davis et al., 1991). Biomolecules reside in an aqueous electrolyte, which affects their conformation and function due to screening and dielectric effects. Aqueous physiological media contain many mobile ions (e.g. Cl-, Na+, K+, Mg++, and Ca++), which redistribute and screen the Coulomb potential of the fixed charges on the macromolecules by creating counter-ion layers. Electrostatic interactions can be shown as a model of electrostatic potential. Use of accurate 3D molecular models is crucial for the determination of electrostatic potential. However, even 3-dimensional crystal structures of macromolecules solved at very high resolutions, do not describe positions of individual atoms in solvent nor the exact chemical composition and ionic state. Most of the solvent molecules around macromolecules are disordered or only partially present. Therefore, it seems that the most appropriate solution for describing electrostatic interactions is the Poisson-Boltzmann approach, which takes into account distribution of solvent molecules (Sharp & Honig 1990). Implicit solvation (or continuum solvation) represents solvent as a continuous medium instead of individual "explicit" solvent molecules (Honig and Nicholls 1995, Still 1990). Empirical force fields such as AMBER (Kollman *et al* 2008), CHARMM (Brooks 1983, 2009) and OPLS (Jorgensen 1988) are widely used in defining the initial charges of macromolecules.

In the presented study, we have investigated the possibilities of using charges obtained from experimental electrostatic potential. Partial charges are not a measurable physical quantity, although this model is useful for the study of the reactivity of certain molecules or for calculating electrostatic potential. The model of atomic point-charges is often used to provide base information about electron shifts in the molecule. The values of partial charges may be obtained directly from quantum mechanical calculations (Mulliken) or by using a fitting procedure which best reproduces the electrostatic potential of the molecule (Singh & Kollman 1984, Bayly 1993). The advantage of fitting is that the electrostatic potential calculated from the electronic density and proton positions is described by a simple model of point charges. The presented results show that fitted partial charges reproduce experimental electrostatic potential well. The RRMS between experimental potential (target) and potential calculated from the partial charges is around 0.35, CC is above 0.90 (valid for all models). The similarity of the potentials is also evident from a comparison of the potential Co_2FoFc and Co_2FoFc-targets in Figures 16 and 17. These results give us the basis for using fitted point charges in further calculations. In the case of small molecules Thym and Heps, the results show that the fitted charges have quite different values compared to Mulliken or semi-empirical charges. The differences between values of individual charges are more than ± 0.5. Most differences were observed in carbon and nitrogen atoms. For example, atom 4N in the Thym molecule (see Table 5) has a value of -0.4 as given by ELBOW, but all fitted charges give a value of about +0.8, while fitted charges of hydrogen and oxygen atoms are consistent with ELBOW charges. As is evident from a comparison of images in Figures 16 and 17, the electrostatic potential obtained from the quantum mechanical or semi-empirical approach is in anti-correlation with the potential calculated from fitted charges. Namely, results in Tables 4 and 5 show that the CC between the semi-empirical or quantum mechanical potential and the potential that was calculated from fitted charges on the basis of experimental data is around -0.50 and -0.8 for Thym and Heps molecules, respectively. The correlation of the potentials based on the experimental electrostatic potential calculated from fitted charges reflect similarity, since all correlations are around 0.8. The advantage of the presented method of partial charges is primarily that the value of the charge is not determined by the type of atom. For example, the C-alpha atom in the protein structure can have positive and negative values. The values of charges are not predetermined as in the standard force field parameters AMBER (Kollman *et al* 2008), CHARMM (Brooks 1983, 2009)). The early force field parameters for amino acids were such that all atoms of the same type have the same charge value. Modern force field parameters use different charge values for the same types of atoms. So the same type of an atom in two different amino acids does not necessarily have the same charge value. However, the restraint that the total charge of a single amino acid is equal to zero (or ± 1 for charged amino acids) is still in use.

It should be noted that fitted partial charges obtained from the experimental electrostatic potential have some limitations. Since the electrostatic potential is fitted in points around the molecule (at vdW surface), the partial charges that are closer to the surface are better defined than the charges in the interior of macromolecules. Therefore, fitting partial point charges without restrictions on the experimental electrostatic potential is just a reproduction of the potential in the vicinity of vdW surface. Moreover, unrestrained fitting leads to physically unreasonable charges. The results in Figure 18 show that for larger molecules such as proteins, convergence of fitting is affected by initial values. In the case of charge fitting against potential St_2FoFc, initial values were set to zero. In this case, the final value of RRMS between

fitted potential and the target potential was 156%. In the second case (St_2FoFc *) when initial charges were set to Mulliken charges, RRMS values (from GAMESS) dropped to 35%. Fitting against St_Fc and St_Fc* showed smaller differences (about 10%). In the case of the small molecules Thym and Heps, the RRMS difference between St_2FoFc and St_2FoFc* was about 5%. It is important to understand that the use of Mulliken charges in the case of small molecules is reflected in higher RRMS, meanwhile in the case of the protein model, the use of Mulliken charges is reflected in lower RRMS.

The electrostatic potential points used for the fitting of partial charges lie close to the van der Waals surface (Singh & Kollman 1984) of the molecule. Hence the estimates of buried charges tend to be poor in comparison with the charges lying close to the surface. The presented model is useful for molecular mechanics as part of the electrostatic force field (Engh & Huber 1991, Parkinson et al. 1996, Andrejašič 2009), and for the necessary input data when solving the Poisson-Boltzmann equation. The quality of the electrostatic potential calculated using the Poisson-Boltzmann method is highly dependent on the quality of input data. This can be paraphrased by the statement "Garbage In, Garbage Out." Even a very good theoretical model cannot give reliable results if the input data is very poor. So the accuracy of the molecular structure (geometry) and accuracy of charge values are essential for the calculation of electrostatic potential.

In the case of the protein model, various sets (AMBER, St_2FoFc* in Co_2FoFc*) of partial charges were introduced as input data for calculating electrostatic potential by the Poisson-Boltzmann method. A comparison between calculated potentials did not show significant differences in the sign (positive / negative) of electrostatic potential. However, significant differences were observed in the "magnitude" of electrostatic potential. Differences in potential magnitude are due to different input data (various set of partial charges). In the case of crambin, fluctuations of fitted main-chain charges St_2FoFc* and Co_2FoFc* can be observed, while charges derived from the AMBER (Kollman *et al* 2008) force field have almost constant values. Fitted partial charges St_2FoFc* and Co_2FoFc* have much higher absolute values than AMBER values. This is reflected in the magnitude of potential. AMBER main-chain charges have values between -0.3 and -0.5, depending on the type of amino acid, while fitted charges occupy values between -3 and +2 independent of the type of amino acid. Due to the use of restraints, each amino acid has a neutral charge (± 1 for charged amino acids). Therefore, main-chain charge value plus side-chain charge value gives zero. This means that each amino acid contains the same number of electrons and protons (except for charged amino acids). So, if the main-chain charge is -2, then side-chain charge must be +2. Partial sums (main and side chain) indicate a very high polarizability of amino acids. This result is far above expectations, as the standard force field parameters predict that the side-chain charge and main-charge are neutral. Fitted charges St_2FoFc* and Co_2FoFc* fluctuate through the entire protein chain independent of the type of amino acids.

The reason for such large fluctuations may be in the WFC method. Namely, WFC was not capable of fully compensating thermal movements of atomic nuclei. However, very similar charge fluctuations were observed when charges were fitted against potential calculated from standard structure factors (St_2FoFc*) obtained directly from SHELX. In both cases (St_2FoFc* and Co_2FoFc*), experimental potential was calculated from experimental electron density (2FoFc map) and from proton “positions”. The 2FoFc electron density map also involves information that does not result in the model, but arises from the experimental data. It follows that the polarization of individual residues is actually much higher and experimentally established.

It seems that the reason for such large differences between the force field charges and fitted charges is that force field charges do not take into account chemical surroundings, while fitted charges through experimental electrostatic potential contain information about the structure of the molecule. Another reason that the electrostatic potential derived charges fluctuate so dramatically is also due to the statistical nature of the fitting process. In the optimization procedure during which charges are derived which reproduce the electrostatic potential around the molecule, the charges on buried atoms can fluctuate wildly in order to yield a miniscule improvement in statistical “quality-of-fit” to the electrostatic potential. The electrostatic potential fitted charges are simply those charges which make an optimal least-square fit of the given model function to a given set of electrostatic potential points. As in any fitting process, some charges will be statistically better determined by the data than others.

For the modeling of electrostatics, both sign and magnitude of the potential are important. The sign gives information about what kind of ligands are capable of interaction with macromolecules, while the magnitude shows the intensity of the potential interaction between two molecules. The magnitude also indirectly indicates how far the influence extends through the solvent. This is important for the understanding of the area (or distance) in which two molecules start to interact with and orient themselves to each other.

6 Conclusions

Implementation of the “averaged kick” map shows that it is possible to improve map quality when compared to the UN or ML maps.

Development and implementation of the WFC method has established a link between the quantum mechanical model of electron density and density obtained by X-ray diffraction of molecules at ultra-high resolution.

Research has shown that at ultra high resolution, the WFC model cannot reproduce the diffraction pattern as well as the standard model because the WFC model currently does not take into account the thermal movements of atomic nuclei. A subject for further development will be introducing the effect of thermal movement in the WFC model. Development of WFC could also be used as an additional restriction by fitting coefficients of quantum mechanical wave function at the experimental electron density.

Results show that polarizability of atoms along the chain of the protein structure is much higher than expected. It is possible that polarizability is due to the inappropriate mathematical model we used, however, the WFC model coincides with the standard model.

Further work will show whether the experimental electrostatic potentials of macromolecules calculated using the WFC method will indeed change the understanding of the magnitude of electrostatic potential and polarizability of amino acids in protein structures, thereby contributing to the understanding of interactions between macromolecules and their ligands.

7 Acknowledgements

Wladek Minor and Maksymilian Chruszcz from Department of Molecular Physiology and Biological Physics, University of Virginia are acknowledged for providing experimental data for molecules Thym and Heps. Milan Hodošček is gratefully acknowledged for help by preparing input files for CHARMM program and for use of computers at Chemistry Institute Ljubljana.

M. Andrejašič and Z. Štefanić are gratefully acknowledged for discussions. The Slovenian Research Agency is acknowledged for funding.

8 References

Adams, P.D. , Grosse-Kunstleve R.W., Hung L.-W., Ioerger T.R., McCoy A.J., Moriarty N.W. ,Read R.J., Sacchettini J.C., Sauter N.K. and Terwilliger T.C. (2002). "*PHENIX: building new software for automated crystallographic structure determination"* Acta Cryst. **D58:** 1948-1954.

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P.D. (2005b). *"A robust bulk-solvent correction and anisotropic scaling procedure." Acta Cryst.* D**61**, 850-855.

Agarwal, R. C. (1978). *"A new least-squares refinement technique based on the fast Fourier transform algorithm."* Acta Cryst. **A34:** 791-809.

Almlöf J. and Taylor P.R. (1991). *"Modern basis sets for quantum chemical calculations."* Adv. Quant. Chem. **22:** 301.

Andrejašič M., Pražnikar J. and Turk D. (2009). *"PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecula structures."* Acta Cryst. **D64:** 1093-1109

Anisimov V.M., Lamoureux G., Vorobyov I.V., Huang N., Roux B., MacKerell A.D., Jr (2005).*"Determination of electrostatic parameters for a polarizable force field based on the classical drude oscillator."* J. Chem. Theory and Comput. ,**1:**153–168.

Baker N. A., Sept D., Holst M. J. and McCammon J. A., (2001). *"The adaptive multilevel finite element solution of the Poisson-Boltzmann equation on massively parallel computers, IBM."* J. Res. Dev., **45:** 427-438.

Brooks B. .R, Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S. and Karplus M. (1983). *"CHARMM: a program for macromolecular energy, minimization and dynamics calculations"* J. Comput. Chem. **4:** 187–217.

Brooks B.R., Brooks C.L. 3rd, Mackerell A.D. Jr, Nilsson L., Petrella R.J., Roux B., Won Y., Archontis G., Bartels C., Boresch S., Caflisch A., Caves L., Cui Q., Dinner A.R., Feig M., Fischer S., Gao J., Hodoscek M., Im W., Kuczera K., Lazaridis T., Ma J., Ovchinnikov V., Paci E., Pastor R.W., Post C.B., Pu J.Z., Schaefer M., Tidor B., Venable R.M., Woodcock H.L., Wu X., Yang W., York D.M., Karplus M. (2009). "*CHARMM: the biomolecular simulation program*", J. Comput. Chem. **30**(10): 1545-614.

Bayly C.I., Cieplak P., Cornell W.D., Kollman P.A. (1993). *"A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges the RESP model."* J. Phys. Chem., **97**:10269–10280.

Bode, B. M. and Gordon, M. S. (1998). *"Macmolplt: a graphical user interface for GAMESS"* J. Mol. Graphics Mod., **16:** 133-138.

Blessing R H, Guo D Y and Langs D A (1998). *"Intensity statistics and normalization"* Direct Methods for Solving Macromolecular Structures, NATO ASI Series Volume, Series C: Mathematical and Physical Sciences vol **507:**47–71.

Bytheway, I., Figgis, B. N. & Sobolev, A. N. (2001). *"Charge density in Cu(glygly)(OH2)(2) center dot H2O at 10 K and the reproducibility of atomic orbital populations" J.* Chem. Soc. Dalton Trans. **22**: 3285-3294.

Cieplak P., Cornell W.D., Bayly C., Kollman P.A. (1995) *"Application of the multimolecule and multiconformational RESP methodology to biopolymers: charge derivation for DNA, RNA, and proteins."* J. Comput. Chem.,**16:** 1357–1377.

Connolly, M. L., (1983). *"Analytical molecular surface calculation"*, Journal of Applied Crystallography, **16:** 548-558.

Connolly, M, L., (1985). *"Molecular surface triangulation."* Journal of Applied Crystallography, **18:** 499-505.

Coppens P. (1997) "*X-ray Charge Densities and Chemical Bonding"* (New York: Oxford University Press)

Coppens, P. (1967). *"Comparative X-Ray and Neutron Diffraction Study of Bonding Effects in s-Triazine.*" Science, **158**, 1577–1579.

Coppens P. (1998). *"X-ray charge densities and chemical bonding"* Acta Cryst. (1998). **B54:** 927

Cornell W.D., Cieplak P., Bayly C.I., Kollman P.A.. (1993) *"Application of RESP charges to calculate conformational energies, hydrogen-bond energies, and free-energies of solvation."* J. Am. Chem. Soc. **115**: 9620–9631.

Cromer D. T. and Mann J. B. (1967). *"Compton Scattering Factors for Spherically Symmetric Free Atoms"* J. Chem. Phys. **47:** 1892.

Dauter Z. , V.S. Lamzin, and K.S. Wilson. (1995). *"Proteins at atomic resolution."* Curr. Op. Struct. Biol., **5**:784-790.

Davidson E. R. , D. Feller (1986), *"A review of the relation between basis set and the accuracy with which properties can be calculated."* Chem. Rev. **86:** 681.

Diercksen G. H. , Wilson S., Eds., Dordrecht, Reidel, *Methods in Computational Molecular Physics,* pp. 71 (1983).

Dolinsky T.J., Czodrowski .P, Li H., Nielsen J.E., Jensen J.H., Klebe G., Baker N.A. (2007) *"PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations."* Nucleic Acids Res, **35:** 522-525.

Dolinsky T.J., Nielsen J.E., McCammon J.A., Baker N.A. (2004). *"PDB2PQR: an automated pipeline for the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations."* Nucleic Acids Res, **32:** 665-667.

Dunning T. H., Hay P. J., *Modern Theoretical Chemistry vol. 4,* H. F. Schaefer III Ed., **1**, Plenum, New York (1977).

Dunning T.H. Jr., K. A. Peterson and D. E. Woon, (1998) *"Correlation consistent basis sets for molecular calculations",* in Encyclopedia of Computational Chemistry (P. von Ragu´ Schleyer, e Ed.), Wiley & Sons, Chichester, UK.

Dunitz J.D. , Schomaker V., and K.N. Trueblood. (1988). *"Interpretation of Atomic Displacement Parameters from Difraction Studies of Crystals."* J.Phys.Chem., **92**: 856-867.

Engh R. A. & Huber R. (1991). *"Accurate bond and angle parameters for X-ray protein structure refinement."* Acta Cryst., **A47:** 392-400.

Fedorov D.G., Ishida T., Kitaura K. (2007) *"FMOutil"*, AIST, Tsukuba, Japan.

Feller D. ,Davidson E. R., (1990) *"Hartree-Fock limit and configuration interaction calculations of the ion-molecule overlap amplitude for hydrogen fluoride: Comparison with EMS experiments."* Rev. Comput. Chem. **1.**

Gilson M. K. , Sharp K. A. and Honig B. H., (1988) *"Calculating the electrostatic potential of molecules in solution - method and error assessment."* J. Comput. Chem., **9:** 327-335.

Gordon M.S., Schmidt M.W. (2005)*"Advances in electronic structure theory: GAMESS a decade later"* 1167-1189, in Theory and Applications of Computational Chemistry: the first forty years C.E. Dykstra, G. Frenking, K.S. Kim, G.E.Scuseria (editors), Elsevier, Amsterdam.

Grimwood, D. J. & Jayatilaka, D. (2001). *"Wavefunctions derived from experiment. II. A wavefunction for oxalic acid dihydrate"* Acta Cryst. **A57**: 87-100.

Guillot B., Viry L., Guillot R., Lecompte C. and Jelsch C. (2001) "*Refinement of proteins at subatomic resolution with MoPro"* J. Appl. Crystallogr. **34** 214–23.

Gunčar, G., Podobnik, M., Pungerčar, J., Štrukelj, B., Turk, V., Turk, D. (1998). *"Crystal structure of porcine cathepsin H determined at 2.1 A resolution: location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function."* Structure, **6:** 51-61.

Gunčar, G., Klemenčič, I., Turk, B., Turk, V., Karaoglanovic-Carmona, A., Juliano, L., Turk, D. (2000). "*Crystal structure of cathepsin X: a flip-flop of the ring of His23 allows carboxy-monopeptidase and carboxy-dipeptidase activity of the protease."* Structure, **8:** 305-313.

Hansen N. K. and Coppens P. (1978) *"Testing aspherical refinement on small molecules data sets"* Acta Crystallogr. **A 34** 909–921.

Herman, F. & Skillman, S. (1962). *"Atomic Structure Calculations "* *Bull. Amer. Phys. Soc.* **7**: 214.

Hohenberg, P. & Kohn, W. (1964). *"Inhomogeneous Electron Gas"* Phys. Rev. B, **136:** 864-871.

Honig B. , Sharp K. and Yang A. S., (1993). "Macroscopic models of aqueous-solutions – biological and chemical applications." J. Phys. Chem., **97:** 1101-1109.

Honig B. and Nicholls A., (1995) *"Classical electrostatics in biology and chemistry."* Science, **268:** 1144-1149.

Housset D., Benabicha F., Pichon-Pesme V., Jelsch C., Maierhofer A., David S., Fontecilla-Camps J. C. and Lecomte C. (2000) *"Towards the charge-density study of proteins: a room-temperature scorpion-toxin structure at 0.96 Å resolution as a first test case"* Acta Crystallogr. D **56:** 151–60.

Humphrey W., Dalke A. , and Schulten K. , (1996). "*VMD-Visual Molecular Dynamics, J. Mol. Graph."* **14**: 33-38.

Huzinaga S., (1985). *"Basis sets for molecular calculations."* Computer Physics reports, Volume 2, **6:** 281-339.

Jayatilaka, D. (1998). *"Wave Function for Beryllium from X-Ray Diffraction Data"* Phys. Rev. Lett. **80:** 798-801.

Jelsch C., Pichon-Pesme V., Lecomte C. and Aubry A. (1998). *"Transferability of multipole charge-density parameters: application to very high resolution oligopeptide and protein structures"* *Acta Crystallogr.* D **54:** 1306– 1308 .

Jelsch C., Teeter M. M., Lamzin V., Pichon-Pesme V., Blessing R. H. and Lecomte C. (2000). *"Accurate protein crystallography at ultrahigh resolution: valence electron distribution in crambin"* Proc. Natl Acad. Sci. USA **97:** 3171–3176.

Jo S., Kim T. , Iyer V.G., and Im W. (2008) *"CHARMM-GUI: A Web-based Graphical User Interface for CHARMM."* J. Comput. Chem. **29**:1859-1865.

Jorgensen W. L. and Tiradorives J. (1988) *"The OPLS potential functions for proteins – energy minimizations for crystals of cyclic-peptides and crambin."* J. Am. Chem. Soc., **110:** 1657- 1666.

Jones T. A., Zou J. Y., Cowan S. W. and Kjeldgaard M. (1991) *"Improved methods for the building of protein models in electron density maps and the location of errors in these models."* *Acta Crystallogr.* A **47:** 110–119.

Kitaura K. , Ikeo E., T. Asada, T. Nakano, M. Uebayasi, (1999)*"Fragment molecular orbital method: an approximate computational method for large molecules"* Chem. Phys. Lett. **313:**701.

Kitaura K. and Morokuma K., (1976). *"A new energy decomposition scheme for molecular interactions within the Hartree-Fock approximation."* Intern.J.Quantum Chem. **10**: 325-340.

Kleywegt, G. J., Henrick, K., Dodson, E. J. & van Aalten, D. M. (2003). *"Pound-wise but penny-foolish: How well do micromolecules fare in macromolecular refinement?"* Structure, **11:** 1051-1059.

Kollman P. A., Case D.A. , Darden T.A., Cheatham T.E. , III, Simmerling C.L. , J. Wang, R.E. Duke, R. Luo, M. Crowley, Ross C. Walker,W. Zhang, K.M. Merz, B.Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F.Wong, F. Paesani, J. Vanicek, X.Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui and V. Babin (2008). "*AMBER 10.*" University of California, San Francisco.

Konnert J. H. (1976). *"A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units"* Acta Cryst. **A32**, 614-617.

Konnert J. H. and Hendrickson W. A. (1980). *"A restrained-parameter thermal-factor refinement procedure"* Acta Crystallogr. A **36:** 344–350.

Koritsanszky, T. S. & Coppens, P. (2001). *"Chemical applications of X-ray charge density analysis"* Chem. Rev. **101:** 1583-1628

Lecomte, C. (1995). *"Experimental electron densities of molecular crystals and calculation of electrostatic properties from high resolution X-ray diffraction."* (In Advances in Molecular Structure Research, **1:** 261-302. ed. by I. and M. Hargittai.) Greenwich, CT, USA: JAI Press, Inc.

Lee, B. and Richards, F.M., (1971). *"The interpretation of protein structures: estimation of static accessibility"*, Journal of Molecular Biology, **55:** 379-400.

Leland L. A. , A. M. Karo, (1960). *"Basis Functions for Ab Initio Calculations"* Rev. Mod. Phys. **32:** 275.

Liu H., Elstner M., Efthimios K., Frauenheim T., Hermans J., Yang W. (2001) *"Quantum mechanics simulation of protein dynamics on long timescale."* Proteins **44;** 484–489.

Lowdin P.O., (1970). *"On the orthogonality problem."* Advances in Quantum Chemistry **5**: 185-199.

Lunin V. Y. & Woolfson M. M. (1993) *"Mean phase error and the map-correlation coefficient"* Acta Cryst. D**49:** 530- 533.

Lunin V.Y. & Skovoroda T. P. (1995)*"R-free likelihood-based estimates of errors for phases calculated from atomic models"* Acta Cryst. A**51:** 880-887.

Minasov F., Wang X. and Shoichet B. K. (2002) *"An ultrahigh resolution structure of TEM-1 beta-lactamase suggests a role for Glu166 as the general base in acylation"* J. Am. Chem. Soc. **124:** 5333–40.

Mitchell K. A. R. (1969). *"Use of outer d orbitals in bonding."* Chem. Rev. **69:** 157-178.

Mulliken R. S. (1955). *"Electronic Population Analysis on LCAO-MO Molecular Wave Functions. III. Effects of Hybridization on Overlap and Gross AO Populations."* J. Chem. Phys. **23:** 2338

Mulliken R. S. (1955) . *"Electronic Population Analysis on LCAO[Single Bond]MO Molecular Wave Functions. I "* J. Chem. Phys. **23:** 1833-1846.

Murshudov G. N.,Vagin A. A., Dodson E. J. (1997). *"Refinement of Macromolecular Structures by the Maximum-Likelihood Method"* Acta Cryst. (1997). D**53:** 240-255.

Muzet N., Guillot B., Jelsch C., Howard E. and Lecomte C. (2003). *"Electrostatic complementarity in an aldose reductase complex from ultrahigh-resolution crystallography and first-principles calculations PNAS"* **100:** 8742–8747.

Nicholls A. and Honig B., (1991) *"A rapid finite-difference algorithm, utilizing successive overrelaxation to solve the Poisson-Boltzmann equation."* J. Comput. Chem., **12:** 435-445.

Nicholls A., Sharp K. A. and Honig B., (1991). "Protein folding and association - insights from the interfacial and thermodynamic properties of hydrocarbons." Proteins, **11:**,281-296.

Otwinowski Z. and Minor W. (1997) *"Processing of x-ray diffraction data collected in oscillation mode"* Methods Enzymol. **276:** 307–326.

Pannu, N. S. and Read, R. J. (1996). *"Improved Structure Refinement Through Maximum Likelihood"* Acta Cryst. A**52:** 659-668.

Parkinson G., Vojtechovsky J., Clowney L., Brünger A. T. and Berman H. M. (1996). *"New parameters for the refinement of nucleic acid-containing structures"* Acta Cryst. **D52**, 57-64.

Pichon-Pesme V., Lecompte C. and Lachekar H. (1995). *"On building a databank of transferable experimental electron density parameters: application to polipeptides"* J. Phys. Chem. **99:** 6242–6250.

Petrova T. and Podjarny A. (2004). *"Protein crystallography at subatomic resolution"* Rep. Prog. Phys. **67:** 1565–1605.

Pražnikar J., Afonine P., Gunčar G., Adams P. and Turk D., (2009). "*Averaged kick maps: less noise, more signal, ... and probably less bias*", Acta Cryst. (has been accepted for publication – DZ5106).

Read, R. J. (1986). *"Improved Fourier coefficients for maps using phases from partial structures with errors"* Acta Cryst. A**42:** 140–149.

Rossmann, M. G. (1972). *"The Molecular Replacement Method."* New York: Gordon & Breach.

Sharp K. A. and Honig B., (1990). *"Electrostatic interactions in macromolecules - theory and applications."* Annu. Rev. Biophys. Biophys. Chem., **19:** , 301-332.

Sharp K., (1991). *"Incorporating solvent and ion screening into molecular-dynamics using the finitedifference Poisson-Boltzmann method"*, J. Comput. Chem., **12:** 454-468.

Sharp K., Jeancharles A. and Honig B. (1992). *"A local dielectric-constant model for solvation free energies which accounts for solute polarizability."* J. Phys. Chem., **96:** 3822-3828.

Schmidt M.W. , Baldridge K.K. ,Boatz J.A. ,Elbert S.T. ,Gordon M.S. ,Jensen J.H. ,Koseki S. ,Matsunaga N. , Nguyen K.A. , S. Su, T.L. Windus, M. Dupuis, J.A. Montgomery (1993). *"General Atomic and Molecular Electronic Structure System"* J. Comput. Chem., **14**,:1347-1363.

Schmidt A., Jelsch C., Østergaard P., Rypniewski W. and Lamzin V. S. (2003) *"Trypsin revisited: crystallography at (sub)atomic resolution and quantum chemistry revealing details of catalysis"* J. Biol. Chem. **278:** 43357–62

Schmidt A. and Lamzin V. S. (2002). *"Veni, vidi, vici—atomic resolution unravelling the mysteries of protein function"* Curr. Opin. Struct. Biol. **12** 698–703.

Sheldrick G.M. (1990). *"Phase Annealing in SHELX-90: direct methods for larger structures."* Acta Cryst. **A46**: 467-473.

Sheldrick G.M. (1993) *"SHELXL-93."* University of Gottingen, Germany.

Sheldrick, G. M. & Schneider, T. (1997). *"SHELXL: High-Resolution Refinement"* in Methods in Enzymology, Vol. **276:** Macromolecular Crystallography, Part B, edited by C. W. Carter Jr & R. M. Sweet, pp. 319–343. New York: Academic Press.

Singh U.C., Kollman P.A. (1984). *"An approach to computing electrostatic charges for molecules."* J. Comput. Chem. **5**: 129–145.

Slater, J. C. (1951). *"A Simplification of the Hartree-Fock Method"* Phys. Rev. **81:** 385-390.

Sordo T. L., C. Barrientos, J. A. Sordo, (1992). *"Computational Chemistry: Structure, Interactions and Reactivity"* S. Fraga, Ed., 23, Elsevier, Amsterdam.

Souhassou M. and Blessing R. H. (1999). *"Topological analysis of experimental electron densities"* J. Appl. Cryst. **32:** 210–217.

Spackman M. A. (1992). *"Molecular electric moments from x-ray diffractions data."* Chem. Rev. **92:** 1769–1797.

Spackman M. A. (1998). *"Charge densities from x-ray diffraction data"* Ann. Rep. Prog. Chem. C: Phys. Chem. **94:** 177–207 .

Stewart, R. F. (1977). *"One-electron density functions and many-centered finite multipole expansions"* Israel J. Chem, **16:** 137-143.

Stewart, R. F. & Feil, D. (1980). *"A theoretical study of elastic X-ray scattering."* Acta Cryst. **A36:** 503-509.

Swinnen M., Thomas A., Brasseur R. (2002). *"FCPAC: fast calculation of partial atomic charges using strictly localised molecular orbitals."* Comput Mater Sci **25:4:** 590–595.

Ten Eyck, L. F. (1973). *"Crystallographic fast Fourier transforms"* Acta Cryst. **A29**:183-191.

Thomas J. R. , B. J. DeLeeuw, G. Vacek, T. D. Crawford, Y. Yamaguchi, and H. F. Schaefer, (1993). *"The Balance Between Theoretical Method and Basis Set Quality: A Systematic Study of Equilibrium Geometries, Dipole Moments, Harmonic Vibrational Frequencies, and Infrared Intensities"* J. Chem. Phys. **99:** 403.

Turk, D., Ph.D. Thesis, (1992), Technische Universitaet, Muenchen, Germany.

Turk, D. (2007). "*Evolving Methods for Macromolecular Crystallography."*, edited by Read, R. J. and Sussman, J. L., 111-122. Springer.

Urzhumtsev A. G., Skovoroda T. P. and Lunin V. Y., (1996). *"A procedure compatible with X-PLOR for the calculation of electron-density maps weighted using an R-free-likelihood approach."* J. Appl. Cryst., **29:** 741-744.

Weeks J. D. , A. Hazi, S. A. Rice, *Advances in Chemistry Physics,* 16, 283 (1969).

Weiner P. K. and P. A. Kollman, (1981). *"Amber - assisted model-building with energy refinement – a general program for modeling molecules and their interactions"* J. Comput. Chem., **2:** 287-303.

Wilson S., D. Moncrieff (1997). *"Distributed Gaussian Basis Sets: Some Recent Results and Prospects."* Advances in Quantum Chemistry, **28:** 47-63.

Woods R.J., Chappelle R.. (2000). *"Restrained electrostatic potential atomic partial charges for condensed-phase simulations of carbohydrates."* J. Mol. Struct. (THEOCHEM) **527**:149–156.

Index of Figures

Index of Tables

Appendix

Appendix A

Fortran code (Wave function coefficients-WFC)

We show simple wave function coefficients for water molecule (H_2O). Wave function coefficients c_1, c_2, c_3, describing electron density around first hydrogen atom, coefficients from c_4 to c_{17} describing electron density around oxygen atom and coefficients c_{18}, c_{19}, c_{20} describing electron destiny around second hydrogen atom.

$$\phi = c_1\varphi(1S) + c_2\varphi(2S) + c_3\varphi(3S) + c_4\varphi(1S) + c_5\varphi(2S) + c_6\varphi(3S) + c_7\varphi(4S) + c_8\varphi(5S) + c_9\varphi(6px) + c_{10}\varphi(6py) + c_{11}\varphi(6pz) + c_{12}\varphi(7px) + c_{13}\varphi(7py) + c_{14}\varphi(7pz) + c_{15}\varphi(8px) + c_{16}\varphi(8py) + c_{17}\varphi(8pz) + c_{18}\varphi(1S) + c_{19}\varphi(2S) + c_{20}\varphi(3S)$$

d(x), d(y), d(z) are representing a distance in x, y, z direction between center of atom and grid point in real space, *r2* is a square of distance:

$r2 = d(x)\,d(x) + d(y)\,d(y) + d(z)\,d(z)$ (see Fortran code below) .

With blue color are signed coefficients from c_1 to c_{20} (see equation abowe). Each contracted basis function is multiplied with corresponding wave function coefficient c_i. Contracted functions showed below are the same as used in quantum mechanical calculations (TZV basic set).

```
        r2=  d(x)**2 + d(y)**2 + d(z)**2
ccccccccccccccccccccc  hydrogen-1
C       1S
        value_in_grid_point =  C(1) *
   $   ( 0.253740E-01 * exp(-74.690 * r2)   +
   $     0.1896840    * exp(-11.230 * r2)   +
   $     0.8529330    * exp(-2.5460 * r2) ) +
C       2S
   $         C(2) *
   $    ( exp(-0.7130 * r2) ) +
C       3S
   $         C(3) *
   $    ( exp(-0.22490 * r2) ) +
ccccccccccccccccccc  carbon
C     1S
           C(4) *
   $         ( 0.7760E-03  * exp(-9471.0 * r2)   +
   $           0.62180E-02 * exp(-1398.0 * r2)   +
   $           0.335750E-01* exp(-307.50 * r2)   +
   $           0.1342780   * exp(-84.540 * r2)   +
   $           0.3936680   * exp(-26.910 * r2)   +
   $           0.5441690   * exp(-9.4090 * r2) ) +
C     2S
   $         C(5) *
   $         (  0.2480750 * exp(-9.4090 * r2)    +
```

```
     $          0.7828440 * exp(-3.5000 * r2) )  +
C    3S
     $        C(6) *
     $        (  exp(-1.0680 * r2) )    +
C    4S
     $        C(7) *
     $        (  exp(-0.40020 * r2) )    +
C    5S
     $        C(8) *
     $        (  exp(-0.13510 * r2) )    +
C    6PX
     $       C(9) * d(x) *
     $        ( 0.162950E-01 * exp(-25.3700 * r2)    +
     $          0.1020980    * exp(-5.77600 * r2)    +
     $          0.3402280    * exp(-1.78700 * r2)    +
     $          0.6682690    * exp(-0.65770 * r2) )  +
C    6PY
     $       C(10) * d(y) *
     $        ( 0.162950E-01 * exp(-25.3700 * r2)    +
     $          0.1020980    * exp(-5.77600 * r2)    +
     $          0.3402280    * exp(-1.78700 * r2)    +
     $          0.6682690    * exp(-0.65770 * r2) )  +
C    6PZ
     $       C(11) *  d(z) *
     $        ( 0.162950E-01 * exp(-25.3700 * r2)    +
     $          0.1020980    * exp(-5.77600 * r2)    +
     $          0.3402280    * exp(-1.78700 * r2)    +
     $          0.6682690    * exp(-0.65770 * r2) )  +
C    7PX
     $       C(12) *
     $        ( d(x) * exp(-0.2480* r2) )  +
C    7PY
     $       C(13) *
     $        ( d(y) * exp(-0.2480* r2) )  +
C    7PZ
     $       C(14) *
     $        ( d(z) * exp(-0.2480* r2) )  +
C    8PX
     $       C(15) *
     $        ( d(x) * exp(-0.91060E-01* r2) )  +
C    8PY
     $       C(16) *
     $        ( d(y) * exp(-0.91060E-01* r2) )  +
C    8PZ
     $       C(17) *
     $        ( d(z) * exp(-0.91060E-01* r2) )  +
ccccccccccccccccccccc  hydrogen-2
 C      1S
        C(18) *
     $  ( 0.253740E-01 * exp(-74.690 * r2)   +
     $    0.1896840    * exp(-11.230 * r2)   +
     $    0.8529330    * exp(-2.5460 * r2) ) +
C      2S
     $       C(19) *
     $   ( exp(-0.7130 * r2) ) +
C      3S
     $       C(20) *
     $   ( exp(-0.22490 * r2) )
```

Appendix B

List of publications related to this Thesis

Andrejašič M., Pražnikar J. and Turk D. (2009). *"PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures."* Acta Cryst. **D64:** 1093-1109.

Pražnikar J., Afonine P., Gunčar G., Adams P. and Turk D., (2009). "*Averaged kick maps: less noise, more signal, ... and probably less bias*." Acta Cryst. (has been accepted for publication – DZ5106).