Linear Scaling Approaches to Quantum Macromolecular Similarity: Evaluating the Similarity Function

PERE CONSTANS
Department of Chemistry, Rice University, Houston, Texas 77005-1892

Received 19 February 2002; accepted 21 May 2002

Abstract: The evaluation of the electron density based similarity function scales quadratically with respect to the size of the molecules for simplified, atomic shell densities. Due to the exponential decay of the function's atom-atom terms most interatomic contributions are numerically negligible on large systems. An improved algorithm for the evaluation of the Quantum Molecular Similarity function is presented. This procedure identifies all non-negligible terms without computing unnecessary interatomic squared distances, thus effectively turning to linear scaling the similarity evaluation. Presented also is a minimalist dynamic electron density model. Approximate, single shell densities together with the proposed algorithm facilitate fast electron density based alignments on macromolecules.

 J Comput Chem 23: 1305–1313, 2002

Key words: molecular similarity; linear scaling; protein alignment; electron density superposition

Introduction

The comparison and classification of macromolecular structures are prior steps to infer the principles that relate structure and function.[1–7] Homology and function prediction studies depend on the assessment and quantification of structural similarity. In essence structural similarity approaches rely on measuring a distance criterion. Structures are striped to a set of relevant elements. Pairs of contacts within inter- or intrastructure elements are then compared. Early approaches quantified similarity through the root-mean-square deviation (RMSD) criterion between interstructural, equivalent atomic positions.[8,9] Similarity was measured after aligning the two compared structures.

Behind the appealing simplicity of RMSD alignments is a factorial search for the meaningful atom-atom equivalences.[10] To break down this factorial scaling several approaches have been devised. Many of the proposed methods can be found in the recent reviews in refs. 11–13. Divide-and-conquer strategies[14] or dynamic programming algorithms based on evolutionary models[15] provide structure alignment approaches in polynomial time. Concretely, for two molecules A and B having N_A and N_B structure elements, respectively, the computational costs appear reduced to $O(N_A^2 N_B^2)$. These, or in fact more expeditive heuristics scaling as $O(N_A N_B)$, produce more reliable solutions the closer in sequence and structure the two molecules are.

The difficulties and discrepancies of these methodologies in recognizing distant homologies,[16–19] as well as the assessment of the significance and appropriateness of several possible similarity scores, are spurring the present interest in macromolecular alignments. The latter factor in the recent interest stems directly from the similarity concept itself. Similarity is a cognitive abstraction. Similarity measurement is a discrimination among the measurements between a given set of physical observables. When inferring from apparently disconnected or poorly perceived phenomena, the significance of a similarity score may be skewed because the relevance of the considered observables is unknown at this stage.

From a perspective of a theory on Quantum Molecular Similarity (QMS), the structure elements are fuzzy, continuous probability density distributions.[20,21] Molecular structure and molecular shape are intrinsically related to the topology of the electron probability density function.[22] Nonetheless, the electron density has emerged as the fundamental observable in chemistry.[23,24] Rigorous electron density approaches scaling linearly with the size of the system have already been described and applied to macromolecular shape similarity analysis.[21,25–29]

Structural measures of similarity among quantum objects are naturally defined as the maximal projections of their attached electron distributions, that is, the maxima of the similarity function. Maximization with respect to the relative displacement and orientation of two molecules leads to a theoretically justified set of alignments. The set of maxima maps onto a corresponding set of common patterns or motifs between the compared structures. Interestingly, a systematic, global search algorithm, polynomial in time, has been established.[30] The intimate structure of the similarity function appears unveiled upon collapsing the molecular electron densities into Dirac delta functions centered at each nu-

Correspondence to: P. Constans; e-mail: constans@ruf.rice.edu

clei. At the peaked limit the similarity function vanishes everywhere, except on a countable subset of points. The total number of maxima appears then bounded by $O(N_A^3 N_B^3)$, where hereafter N_A and N_B are the number of atoms in the molecules A and B, respectively. Such an upper bound also indicates the maximal complexity of the global search procedure itself. Smoother molecular fields lower the actual global search complexity. Thus, for noncollapsed electron densities a reliable algorithm scaling $O(N_A N_B)$, or in short $O(N^2)$, has been devised and extensively tested.[30] This scaling cost is referred to the required similarity function evaluations. Each evaluation, in turn, implies $O(N^2)$ integrals when using approximate electron densities. Due to the exponential decay of the density overlaps as a function of the interatomic distances, the number of non-negligible integrals scales only linearly on large systems.

Here it is shown that overlap integrals are accurately approximated by using a minimalist dynamic electron density model. The overall cost of the macromolecular alignment is then dominated by the $O(N^2)$ distance evaluation among the density centers. To avoid unnecessary distance evaluations a new algorithm is presented. It permits the computation of the similarity function as a true linear scaling procedure. The proposed methodology expands the established techniques of QMS to macromolecules. This work is aimed to facilitate purely structural alignments and measures of similarity, that is, similarities not biased from the outset by equivalences obtained from sequence alignment procedures. The introduction of lower scaling search strategies and the assessment of the QMS scores as meaningful measures are left to separate works.

The article is organized as follows. The section Macromolecular Similarity introduces the QMS measures. A minimalist, dynamic density model for macromolecules is proposed and its accuracy analyzed in terms of atom-atom similarity potentials. Then a linear scaling algorithm for the evaluation of the QMS function is presented and described in detail. In the section Benchmark Computations these procedures are tested on several sets of protein structures.

Macromolecular Similarity

Quantum Similarity Measures

In quantum chemistry, molecular structure may be regarded as the expectation values for the positions of the constituting particles, the nuclei, and the electrons. The density of probability $\rho(\mathbf{X})$ for an arrangement $\mathbf{X}$ of the space and spin coordinates of the particles is thus the modulus of the attached wave function. Further chemical insight is gained from the density $\rho(\mathbf{X})$ after decoupling electronic and nuclear motions, through the Born and Oppenheimer approximation[31]:

$$\rho(\mathbf{X}) = \rho_e(\mathbf{X}_e; \mathbf{X}_n) p_n(\mathbf{X}_n) \tag{1}$$

Here the electronic structure $\rho_e(\mathbf{X}_e; \mathbf{X}_n)$ is assumed to instantly readapt to a given, fixed arrangement of the nuclei. In turn nuclear motions, if considered, are computed through an effective, electron-averaged Hamiltonian. According to this partitioning the nuclear arrangements $\mathbf{X}_n$ are directly related to conformational changes and chemical reactivity. On the other hand the electron distributions $\rho_e(\mathbf{X}_e; \mathbf{X}_n)$ contain the necessary information to determine the properties of a molecule at the conformation $\mathbf{X}_n$. In fact, solely the information consisting of the number of electrons per volume unit, corresponding to the integration of ρ_e with respect to all spatial coordinates but one, and to all spin coordinates, is required for most molecular properties. This leads to a simpler or reduced function, $\rho_e(\mathbf{r}; \mathbf{R}_n)$, or in short, $\rho(\mathbf{r})$. The observable charge density $\rho(\mathbf{r})$ is a real-valued function on a three-dimensional Euclidian space, thus naturally identifiable with our common structural space.

The analysis and comparison of the electronic distributions constitute the essence and the foundation of the measures of quantum molecular similarity. The similarity measures between two molecules A and B are identified with the maxima of the similarity function with respect to the set of mutual displacements and orienting angles Ω. In turn, the similarity function is defined as a generalized projection of the density functions $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$:

$$z_{AB}(\Omega; \Theta) = \iint \rho_A(\mathbf{r}_1; \mathbf{R}_A)\Theta(\mathbf{r}_1, \mathbf{r}_2)\rho_B(\mathbf{r}_2; \mathbf{R}_B(\Omega))d\mathbf{r}_1 d\mathbf{r}_2 \tag{2}$$

Here, and without loss of generality, only the molecular coordinates $\mathbf{R}_B$ are translated and rotated. By including the positive definite operator $\Theta(\mathbf{r}_1, \mathbf{r}_2)$, a family of density derived similarity functions is embraced within this single formalism. Preferences on a particular similarity function are dictated by the specificity of its application. Simple structural measures use the Dirac delta operator $\delta(\mathbf{r}_1 - \mathbf{r}_2)$. Hereafter, only these overlap measures are considered. The indices of similarity are normalized measures, that is

$$C_{AB} = z_{AB}(z_{AA}z_{BB})^{-1/2} \tag{3}$$

The Carbó index defined above takes a value of one for identical structures and tends to zero the more different the two structures are.[32]

Practical evaluation of the QMS function benefits from the use of simplified density models. The Atomic Shell Approximation (ASA) densities present the general form

$$\rho_{ASA}(\mathbf{r}) = \sum_a \sum_{i \in a} n_i c_i e^{-\xi_i(\mathbf{R}_a - \mathbf{r})^2} \tag{4}$$

with the shell occupations n_i constrained to positive values.[33] With respect to *ab initio* densities, the ASA permits a scaling reduction in the evaluation of the similarity function from $O(N^4)$ to $O(N^2)$. The more expensive four-center overlap integrals appear substituted by isotropic, two-center overlaps. Analytical gradients and Hessians with respect to translations and rotating angles are also strongly simplified.

The impact of this isotropic, pseudoatomic constraint on the resulting overlap QMS measure is reduced. Core densities remain mainly unaffected upon bond formation. The departures from the atomic sphericity occur at comparatively low density regions. Thus the asphericity of the atoms in the molecules scarcely contributes

to the overlap integral value. An analogous low contribution is also shown by the interatomic charge transfer. This fact permits modeling of the electron densities simply as promolecular ASA expansions. The accuracy of these approaches when compared to fully *ab initio* densities has been analyzed in refs. 33 and 34.

Macromolecular Electron Density Representation

Promolecular ASA models are effective for structural similarity measurements. The QMS function, though nearly unaffected by chemical electron rearrangements, is sensitive to the atomic positioning. Neat, clamped nuclei structure idealizations reinforce this sensitivity due to their implicit exactitude. Proteins and related biomolecules, however, are essentially dynamic systems. Fluctuations up to 2 Å for side residues are not rare in dynamic modeling studies. X-ray protein structures present very high mean-square displacement parameters. Mobile chain terminal regions may indeed not be observed in the Fourier electron density maps.[35]

Accessible, crystallographic macromolecular structures are fuzzy distributions around the preferential nuclei positions $\bar{\mathbf{R}}_n$. The usual one particle potential model[36] approximates the nuclear distribution to

$$p_n(\mathbf{R}_n) = p_1(\mathbf{u}_1)p_2(\mathbf{u}_2)\cdots p_N(\mathbf{u}_N) \tag{5}$$

where, for each nucleus a, the displacement $\mathbf{u}_a$ is the difference $\bar{\mathbf{R}}_a - \mathbf{R}_a$. By further assuming that nuclei individually move as isotropic and harmonic oscillators, their probability density function follows the Gaussian distribution

$$p_a(\mathbf{u}_a) = \frac{e^{-u_a^2/2\langle u_a^2\rangle}}{(2\pi\langle u_a^2\rangle)^{3/2}} \tag{6}$$

The mean-square displacement $\langle u_a^2\rangle$ is normally expressed in terms of the temperature parameter B, being $\langle u_a^2\rangle = B_a/8\pi$.

The resulting, dynamic electron density $\tilde{\rho}(\mathbf{r})$ follows the convolution of the electronic and nuclear distributions[37]:

$$\tilde{\rho}(\mathbf{r}) = \rho(\mathbf{r}; \mathbf{R}_n) * p_n(\mathbf{R}_n) \tag{7}$$

Within the ASA, dynamic densities read

$$\tilde{\rho}_{\mathrm{ASA}} = \sum_a \tilde{\rho}_a(\mathbf{r}) \tag{8}$$

with the pseudoatom contributions $\tilde{\rho}_a$ given by the convolution

$$\tilde{\rho}_a(\mathbf{r}) = \int \rho_a(\mathbf{R}_a - \mathbf{r})p_a(\bar{\mathbf{R}}_a - \mathbf{R}_a)d\mathbf{R}_a \tag{9}$$

Upon integration, eq. (9) is formally analogous to the previous eq. (4). Shell occupations remain invariant, but the shape of the shells appears smeared. The new shell exponents depend on the nuclear delocalization, given by

$$\tilde{\zeta}_i = \frac{\zeta_i}{2\langle u_a^2\rangle\zeta_i + 1} \tag{10}$$

The larger the nuclear delocalization, the more the electron distribution settles into a single shell with its exponent tending to $1/2\langle u_a^2\rangle$. This collapse of the shell structure provides an effective simplification for the macromolecular electron densities.

Atomic, single shell densities $\bar{\rho}_a$ are simply

$$\bar{\rho}_a(\mathbf{r}) = n_a\left(\frac{\zeta_a}{\pi}\right)^{3/2} e^{-\zeta_a(\bar{\mathbf{R}}_a - \mathbf{r})^2} \tag{11}$$

with just one single parameter ζ_a to be adjusted. From several possible choices, constraining $\bar{\rho}_a$ to reproduce the density expectation value $\langle\tilde{\rho}_a|\tilde{\rho}_a\rangle$, that is, taking ζ_a as

$$\zeta_a = 2\pi n_a^{-4/3}\langle\tilde{\rho}_a|\tilde{\rho}_a\rangle^{3/2} \tag{12}$$

constitutes an advantageous approach. These single shell densities reasonably represent the dynamic electron densities within the usual range of nuclear displacements. Figure 1a and 1b compares the atomic densities $\tilde{\rho}_a$ to the single shell approach $\bar{\rho}_a$ for the carbon and oxygen atoms, respectively. These dynamic densities consider a typical nuclear displacement of 0.62 Å or B parameter of 30 Å^2.

The overlap QMS function becomes, after substituting the single shell densities into eq. (2), a double sum of isotropic interatomic contributions:

$$Z_{\mathrm{AB}}(\Omega) = \sum_a \sum_b z_{ab}(r_{ab}(\Omega)) \tag{13}$$

The terms $z_{ab}(r_{ab})$ are computed simply as

$$z_{ab}(r_{ab}) = n_a n_b\left(\frac{\mu_{ab}}{\pi}\right)^{3/2} e^{-\mu_{ab}r_{ab}^2} \tag{14}$$

with μ_{ab} being $\zeta_a\zeta_b/(\zeta_a + \zeta_b)$. Interestingly, when the single shells exponents are taken from eq. (12) the overlaps z_{ab} are accurate. Figure 1c plots the ASA and single shell z_{CO} functions for the carbon-oxygen overlap. Thus, such a choice in the exponents can be regarded as a successful parameterization of these atom-atom similarity potentials.

Linear Scaling Algorithm for Macromolecular Similarity Evaluation

Even though important simplifications are introduced by minimalist densities in the QMS computations, the evaluation of $O(N^2)$ terms in eq. (13) prevents large molecule comparisons. Most of those terms are, however, numerically negligible due to the exponential decay of the density overlaps. Usual scaling reduction techniques exploit the locality of most physical interactions to expand the range of computability to larger systems.[38–40] An a priori knowledge on whether or not an interaction falls below a given accuracy threshold permits substantial reductions indeed in the scaling of the computation.

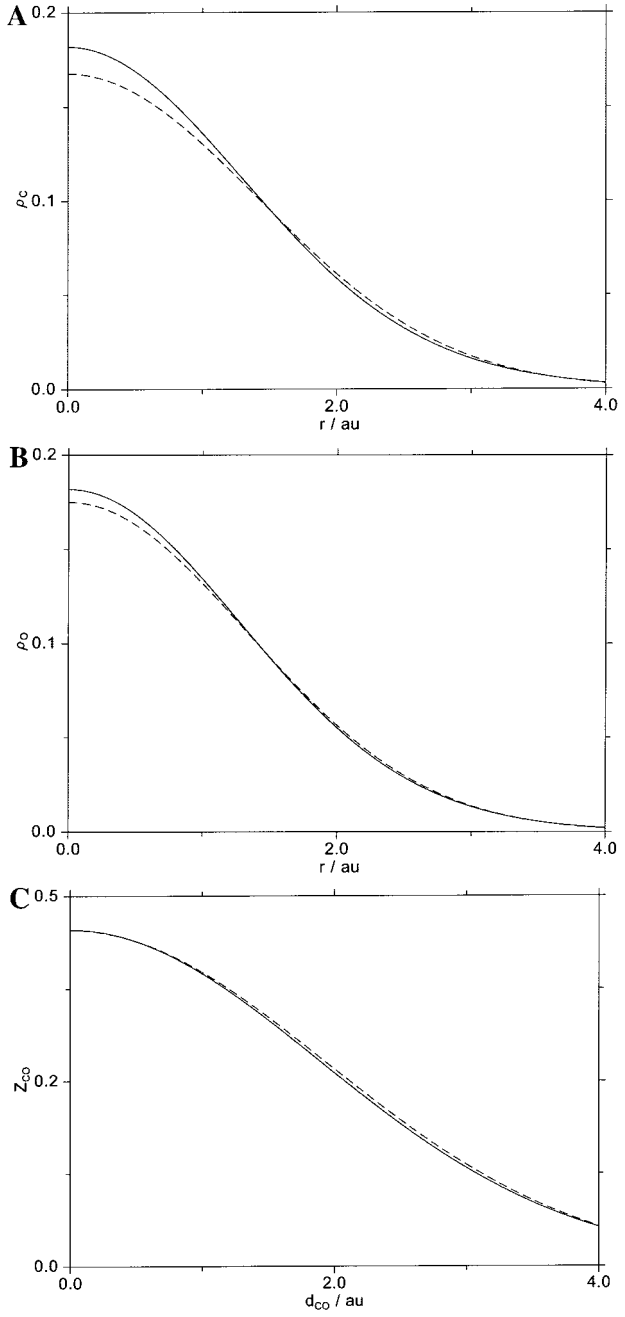

Figure 1. (a) Carbon and (b) oxygen dynamic densities. (c) Carbon-oxygen density overlap z_{CO}. Solid lines are the dynamic ASA while dashed lines are the single shell approach. Parameter B = 30 Å^2.

Accordingly, the QMS function in eq. (13) should be rewritten and evaluated as

$$Z_{AB}(\Omega) = \sum_{b \in J} \sum_{a \in I_b} z_{ab}(r_{ab}(\Omega)) \tag{15}$$

where the sets I and J index for all elements in molecules A and B, respectively, and the subset $I_b \subset I$ indexes only for non-negligible terms:

$$I_b = \{a | r_{ab} < r_{cut}\} \tag{16}$$

Because the number of elements in lists I_b is not size extensive, the computation of Z_{AB} scales linearly provided that lists I_b is determined in an $O(1)$ procedure.

The proposed algorithm, which uses a so-called cell-link list technique,[41] computes Z_{AB} through eq. (15) according to the following steps:

1. **Initialize molecules.** Read atomic coordinates, atomic numbers, and temperature parameters. Orientate molecules to the *standard* orientation. Construct promolecular dynamic ASA densities.
2. **Set cutoff radius from accuracy threshold.** Given an accuracy threshold ε, determine a r_{cut}: $z_{\mu\nu}(r_{cut}) < \varepsilon$, $\forall \mu, \nu \in I \cup J$.
3. **Discretize.** Set $l \leftarrow r_{cut}$. Inscribe molecule A in a cuboid of edge lengths $n_x l$, $n_y l$, and $n_y l$. Consider for each cube of edge length l inside cuboid A an identification 3-tuple (i, j, k).
4. **Construct near-neighbor lists.** $\forall a \in I$ identify its container cube (i, j, k). Sequentially construct an occupied cube list $\mathcal{L}$: $(i, j, k) \rightarrow n$ and a near-neighbor list $\mathcal{N}$: $n \rightarrow \{a_{ijk}\}$, where $\{a_{ijk}\}$ is the set of atoms in a neighborhood of cube (i, j, k).
5. **Evaluate Z_{AB}.** $\forall b \in J$ identify its container cube (i, j, k) in A. Evaluate $\sum_{a \in I_b} z_{ab}(r_{ab})$ where I_b comes through the double list $(i, j, k) \overset{\mathcal{L}}{\rightarrow} n \overset{\mathcal{N}}{\rightarrow} \{a_{ijk}\}$.

The first step initializes data. Original coordinates are oriented according to the charge tensor principal axes. This minimizes the required number of boxes and, therefore, the memory requirements for list $\mathcal{L}$.

The linear scaling issue in step 2 merits some attention. The set of matrices $\mathbf{z}_r = \{ z_{\mu\nu}(r) \}$ may be indefinite for some particular r, ρ_μ, and ρ_ν. Therefore, in general, Cauchy-Schwarz inequality $z_{\mu\nu}(r) \leq \max(z_{\mu\mu}(r), z_{\nu\nu}(r))$ will not apply and other inequalities must be used to determine an advantageous r_{cut} in only $O(N)$ steps. A possibility is to compute r_{cut} through the *shifted* rule:

$$r_{cut} = \sup_{\mu \in I \cup J} \left\{ \frac{-2 \log \varepsilon - 3 \log(\zeta_\mu + \zeta_{min}) + 2 \log c_\mu c_{max} + 3 \log \pi}{\zeta_\mu} \right\} \tag{17}$$

Quantities ζ_{min} and c_{max} are respectively the minimal exponent and the maximal coefficient. Yet, the optimal, nonconservative r_{cut} can be obtained in $O(1)$ step in the particular case of equal atomic displacements. In this case, the number of different atomic densities is in fact the number of different chemical elements.

After establishing an appropriate cutoff distance, the next step, 3, divides the cuboid that inscribes the fixed molecule A into small cubes of edge length equal to r_{cut}. Each cube is identified by the

Cartesian indices (i, j, k) according to the positioning of its origin in the X, Y, and Z axis, respectively. In addition, for each cube, a counter that sums up the cube population is initialized.

In the following step, 4, the two lists, $\mathcal{L}$ and $\mathcal{N}$, are sequentially constructed. For each atom a in molecule A, the cube containing a is identified. The cube population counter is increased by one, as well as the 26 near-neighbor counters. Each time a cube is populated the occupied-cube counter n is increased. The list $\mathcal{L}$ maps the Cartesian indices of occupied cubes to their ordinals, that is, $\mathcal{L}_{ijk} \rightarrow n$. The ordinal of atom a is added to the near-neighbor list $\mathcal{N}$ at the cube entries n and n-adjacents. Once the process is completed for all atoms in A, the entry $\mathcal{L}_{ijk}$ points to the set of near-neighbors at the $\mathcal{N}_n$ entry, if the cube ijk or the adjacent ones are occupied. Otherwise, it points to the null entry. Memory and CPU requirements for the construction of the $\mathcal{L}$ and $\mathcal{N}$ lists scale linearly. This construction is a prior, separate step in the evaluation of the similarity function and it is only required once.

The evaluation of the similarity Z_{AB} is conducted at step 5. For each atom b in the moving molecule B, the position indices (i, j, k) at the fixed frame of molecules A are computed. By looking up the entries $\mathcal{L}_{ijk}$ and $\mathcal{N}_n$ it is identified in $O(1)$ operations a set including all atoms a whose distance to b is less than r_{cut}. Consequently, this set contains I_b. By construction, some distances greater than r_{cut} but less than $2\sqrt{3}r_{\mathrm{cut}}$ are also included. The evaluation of Z_{AB} is done at this point using eq. (15). Each of the steps in this sequence, as well as the rotation and translation of molecule B, scale as $O(N)$ at most. Therefore, the whole process of evaluating $Z_{\mathrm{AB}}(\Omega)$ scales linearly.

A pseudocode translating the steps above is introduced in Table 1 to further clarify the algorithm. In addition, it includes an extra cutoff statement when evaluating Z_{AB}. It prunes the set of atoms produced by lists $\mathcal{L}$ and $\mathcal{N}$ to fit exactly the I_b subset. This reduces the computation prefactor and also permits a simpler performance analysis.

Benchmark Computations

Two groups of benchmark computations are presented to illustrate and analyze the performance of the described algorithm. The first one focuses on its linear scaling behavior while the second one focuses on the individual timings for the several parts of the alignment process. Comparisons are made to equivalent calculations where the I_b subsets are simply identified through an $O(N^2)$ squared distance evaluation. The examples consider the complete alignment of several protein structures. The alignment procedure is as follows:

1. **Global maximization.** A subset of 40 atoms, equally spaced in sequence, is selected for each protein. These atoms are used as *directing centers* in the global search procedure Algorithm LA/I in ref. 30. This produces a total of 1521 nonrepeated exploratory translations and rotations. The similarity function is evaluated at each arrangement. The 10 highest ranked arrangements from the total 1521 are selected as starting maximizers to be refined in the local optimization.
2. **Local maximization.** The local maximization is performed according to the Greenstadt modification of Newton's algorithm.[42] Analytical gradients and Hessians are computed at each optimization step. The iteration stops once the modulus of the gradient goes below a 10^{-2} threshold and all the eigenvalues of the Hessian are negative. Optimization is stopped if convergence is not reached within 50 iterations. The optimization is performed for all 10 selected arrangements. Converged solutions are sorted and sifted for unique, nonrepeated alignments.

The protein structures are taken from the Protein Data Bank (PDB).[43,44] All ATOM entries are considered. Promolecular electron densities are constructed through available tables of ASA densities.[45] For the sake of an easier comparison, all atomic mean displacements B are taken equal to 30 Å^2. Prescribed accuracy for the $z_{\mu\nu}$ pairs is set to 10^{-5}. The cutoff distance is obtained from the exact $O(1)$ procedure, being $r_{\mathrm{cut}} = 5.38$ Å. The $O(N)$ rule in eq. (17) would produce here a just slightly more conservative cutoff distance equal to 5.44 Å. All calculations are performed on a PC/Linux platform equipped with a single PII/300 MHz CPU. These algorithms are implemented in the program LSIM, which is available to download.[46]

The first benchmark consists of a sequence of pair alignments of successively enlarged fragments. The two fragments in each pair possess an equal number of atoms. Also, each fragment is extracted from two separate structures. The first structure is the HIV-1 reverse transcriptase (RT) PDB code 1dlo,[47] and the second one is the HIV-1 RT complexed with 8C1-TIBO, PDB code 1uwb.[48] The latter shows some ligand-induced variations with respect to the former, although an overall similarity is approximately maintained throughout the successive fragment enlargements. This facilitates the observation of the scaling behavior of the computation because the timings depend on the extent of similarity. In Figure 2 the total timings for each alignment are plotted as a function of the fragment size. By fitting these values to the power function $t(N) = aN^b$, the determined parameter b is 0.9 for the near-neighbor list algorithm. On the other hand, when all squared distances are evaluated, b is equal to 1.9. At the prescribed accuracy the crossover, that is, the crossing of these two curves, occurs approximately between molecular sizes of 300–350 atoms. These plots clearly show that the $O(N^2)$ squared distance evaluation dominates the overall computation at the macromolecular sizes.

The selected structures for the second benchmark consist of a set of three α-globins and a set of three TIM α/β-barrel proteins, the latter approximately triple the size of the former. On the first set the myoglobin molecule, PDB code 1mbs,[49] is compared to its homologous 1bzp[50] and to hemoglobin structure 1sct, chain A.[51] On the second set the acid α-amylase, PDB code 2aaa,[52] is compared to the TAKA amylase 7taa,[53] and to the glutamate mutase 1cb7,[54] chain B, which belongs to a different family classification.

Detailed timings are reported in Table 2. Timings for the near-neighbor list construction are denoted by t_{NNL} and the averages for molecule translation and rotation by t_{MM}. The similarity function is evaluated through two separate routines whenever function only or function and derivatives are required. The function-only routine is called at the global search stage. The integrated function and derivative evaluation is used at each cycle in the local optimization. The

Table 1. Pseudocode Algorithm for the Linear Scaling Evaluation of the Similarity Function.

Near-neighbor list linear scaling algorithm

Input $A = \bigcup_{a \in I} \rho_a(\mathbf{r}_a - \mathbf{r})$, $I = \{1, 2, \ldots, N_A\}$, $\mathbf{r}_a \leftarrow \mathbf{r}_a^0$

$B = \bigcup_{b \in J} \rho_b(\mathbf{r}_b - \mathbf{r})$, $J = \{1, 2, \ldots, N_B\}$, $\mathbf{r}_b \leftarrow \mathbf{\Omega}_T + \mathbf{\Omega}_R \mathbf{r}_b^0$

ε: $\varepsilon > z_{ab}(r_{ab}) \leftarrow 0$

Output $Z_{AB} = \sum_{b \in J} \sum_{a \in I_b} z_{ab}(r_{ab})$, $I_b = \{a | r_{ab} < r_{cut}, a \in I, b \in J\}$

Near neighbor list

```
Set                          d ← r_cut | z_μν(r_cut) < ε ∀μ, ν ∈ I ∪ J
Initialize                   Lbox ← 0
                             LNN ← 0
                             mbox ← 0
Establish boxes              dxA = xA_max − xA_min
                             dyA = yA_max − yA_min
                             dzA = zA_max − zA_min
                             mbx = 1 + dxA/d
                             mby = 1 + dyA/d
                             mbz = 1 + dzA/d
                             nbox = (mbx + 2)(mby + 2)(mbz + 2)
Resize cuboid origin         xA_min, yA_min, zA_min  to exactly fit d-cubes
Establish populations:
Do loop                      for a = 1 to N_A
                               ix = 1 + (xA(a) − xA_min)/d
                               jy = 1 + (yA(a) − yA_min)/d
                               kz = 1 + (zA(a) − zA_min)/d
Do loop neighbor cubes         for i, j, k, = ix − 1, yj − 1, kz − 1 to ix + 1, jy + 1, kz + 1
                                 ibox = Lbox(i, j, k)
                                 if ibox = 0 then
                                   mbox = mbox + 1
                                   ibox = mbox
                                   Lbox(i, j, k) = ibox
                                   LNN(ibox, 0) = 0
                                 endif
                                 Inn = LNN(ibox, 0) + 1
                                 LNN(ibox, 0) = Inn
                                 LNN(ibox, Inn) = a
                               endfor i, j, k
                             endfor a
Function Z_AB
                             Z_AB = 0
                             for b = 1 to N_B
                               if ZBOX(xB(b), yB(b), zB(b), ibox) then
                                 nter = LNN(ibox, 0)
                                 for i = 1 to nter
                                   a = LNN(ibox, i)
                                   dx = xA(a) − xB(b)
                                   dy = yA(a) − y(B)b
                                   dz = zA(a) − zB(b)
                                   r^2 = dx dx + dy dy + dz dz
                                   if r^2 < d^2 then
                                     Z_AB = Z_AB + z_ab(r)
                                   endif
                                 endfor i
                               endif
                             endfor b
                             end Z_AB
Logical function ZBOX
                             ZBOX = false
                             Ibox = 0
                             i = 1 + (x − xA_min)/d
                             if i < 0 return
                             if i > mbxp return
                             j = 1 + (y − yA_min)/d
                             if j < 0 return
                             if j > mbyp return
                             k = 1 + (z − zA_min)/d
                             if k < 0 return
                             if k > mbzp return
                             ibox = Lbox(i, j, k)
                             ZBOX = true
                             end ZBOX
```

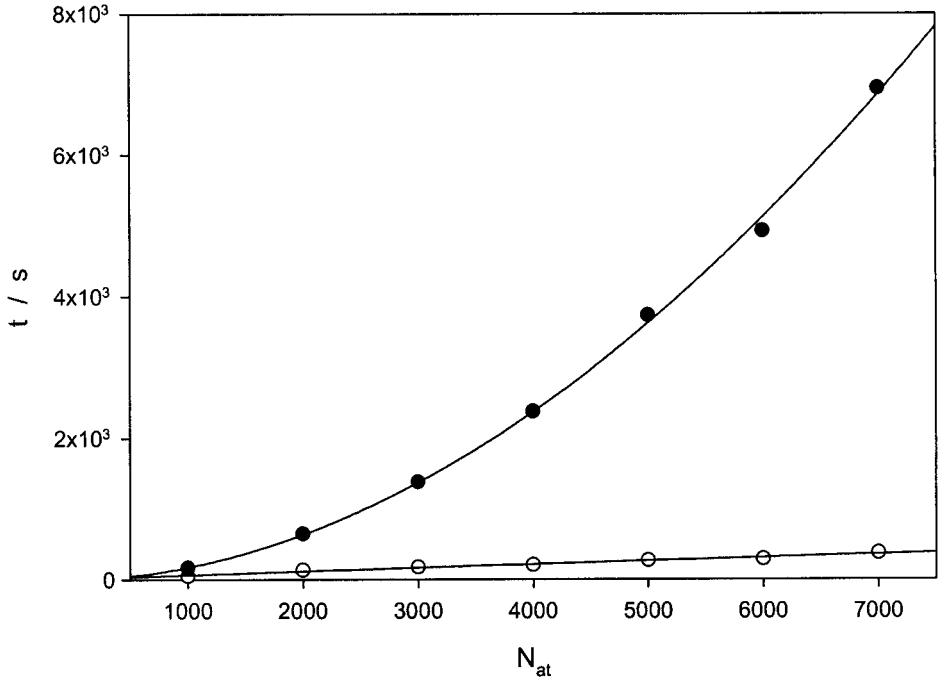

Figure 2. Timings for the alignment of N-atom fragments extracted from the PDB structures 1dlo and 1uwb. Solid marks indicate the time for a quadratic scaling computation, while open marks correspond to the linear scaling algorithm computation. Cutoff radius r_{cut} = 5.38 Å.

quantities t_{Z1} and t_{Z2} stand for the respective averaged timings. Values in italics refer to the timings for the $O(N^2)$ algorithm. In addition, molecule sizes N and sequence and electron density similarities S_{AB} and C_{AB}, respectively, are also included.

The near-neighbor list construction is a fast process, as seen in the t_{NNL} column. Because the lists just depend on the fixed molecule, no further updating is required throughout the alignment process or for multiple molecule comparisons. The timings t_{MM} are negligible as well. The overall cost of the calculations is dominated by the similarity evaluation, see columns t_{Z1} and t_{Z2}. Reductions introduced by avoiding the $O(N^2)$ squared distances are remarkable. For the globins the speed-up is approximately three times, while that for the TIM-barrels is already about an order of magnitude. In passing, it is worth noting that the integrated evaluation of similarity function, gradient, and Hessian is just from two to two and one-half times more expensive than the similarity alone. This ratio is an important issue for the design of effective optimization schemes.

The total timings t_T correspond to the elapsed times for the complete computation. Essentially this includes the 1521-point

Table 2. Detailed Timings for the Alignment of the PDB Structures in Columns A and B.

A	N_A	B	N_B	S_{AB}	C_{AB}	t_{NNL}[a]	t_{MM}[a]	t_{Z1}[a]	t_{Z2}[a]	t_T[b]
1mbs	1223	1bzp	1343	84	0.66	18.1	0.6	43.1	104.5	97.9
								147.6	*214.7*	*286.8*
		1sct	1134	19	0.49	18.3	0.5	38.3	88.1	87.7
								126.0	*181.4*	*248.3*
2aaa	3669	7taa	3688	67	0.84	34.1	2.1	124.6	328.7	288.2
								1048.9	*1348.1*	*1979.8*
		1cb7	3762	4	0.36	33.2	2.3	125.3	289.0	284.5
								1235.2	*1538.0*	*2347.3*

[a]In ms.
[b]In s.
Their respective number of atoms are N_A and N_B, and their sequence and density-based similarity scores are S_{AB} and C_{AB}. Denoted by t_{NNL}, t_{MM}, t_{Z1}, and t_{Z2} are the timings for the near-neighbor list construction, molecule moving, similarity only, and similarity and derivatives evaluation, respectively. Total elapsed time is t_T.

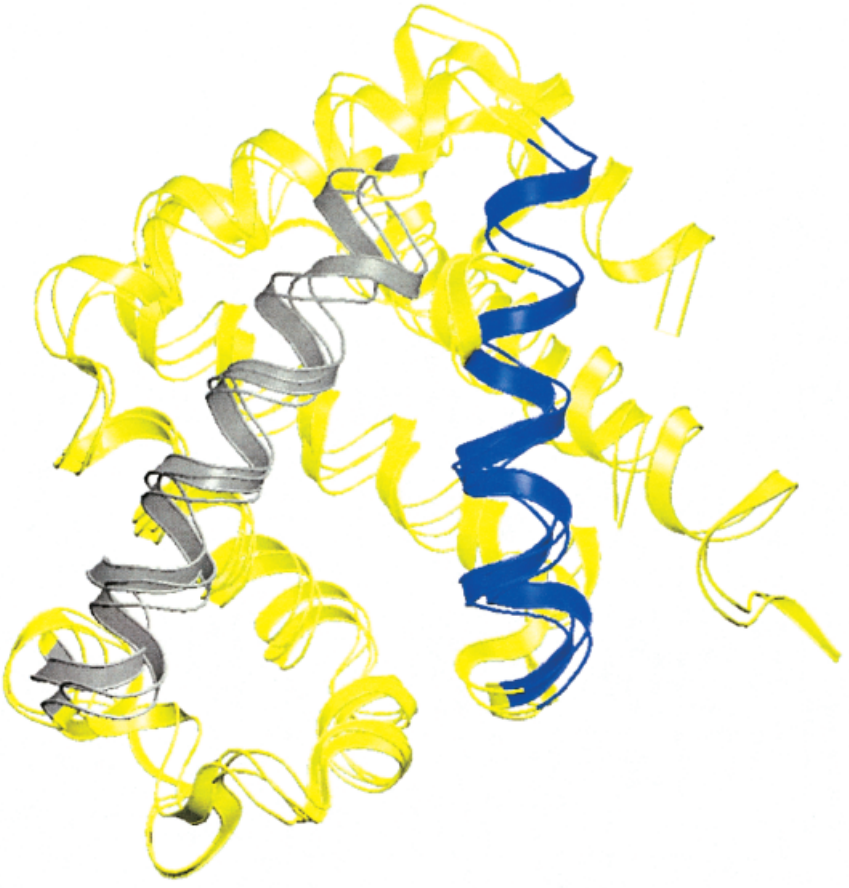

Figure 3. Electron density-based alignment of myoglobin structure 1mbs with hemoglobin 1sct, chain A. The two heme-linked helices E, in dark gray (gray), and helices F, in black (blue), are highlighted. Plotted using VMD[56] and Raster3D.[57]

similarity evaluation in the global search and around 280 iterations to perform all 10 local optimizations. At this level of completeness, an all-atom comparison of two structures ranges from 1 to 5 min for proteins of approximately 150 to 500 residues. The resulting alignments appear already accurate. A paradigmatic example is the alignment of the myoglobin 1mbs with the hemoglobin 1sct. These two proteins present a low, 19% sequence identity although their structural similarity is notorious.[55] Figure 3 plots the electron density-based superposition, comparable to the one shown in ref. 55, Figure 4. Important features regarding structure-function studies, that is, the matching of the two heme-linked E and F helices, are accomplished. Figure 3 is plotted using VMD and Raster3D. Therefore, and foreshadowing attainable improvements in search strategies, these timings should be considered as upper bound for electron density based alignments of proteins.

Conclusions

This work has analyzed the reliability of minimalist electron density models and has introduced an improved, linear scaling algorithm for the evaluation of the QMS function. Reductions in computation are remarkable to the point of permitting protein alignments at a reasonable accuracy and cost. Further nearing the theory on QMS to structural biophysics is quite an endeavor. Comparisons of large molecules lead to a number of significant alignments, each revealing a common pattern with respect to a specific measure. This produces a wealth of structural information to be processed and integrated into evolutionary and functional data. The QMS information possesses a unique, *structure only* basis that may turn out to be adequate and complementary for the clarification of the relationships among structure, evolution, and function.

Acknowledgments

I am in debt to Prof. Scuseria for bringing to my attention the due importance of algorithm scalings. I thank S. Vega for her careful reading of the manuscript.

References

1. Koehl, P. Curr Opin Struct Biol 2001, 11, 348.
2. Lo Conte, L.; Ailey, B.; Hubbard, T. J. P.; Brenner, S. E.; Murzin, A. G.; Chothia, C. Nucleic Acids Res 2000, 28, 257.
3. Brenner, S. E.; Koehl, P.; Levitt, R. Nucleic Acids Res 2000, 28, 254.
4. Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2001, 29, 228.
5. Bray, J. E.; Todd, A. E.; Pearl, F. M. G.; Thornton, J. M.; Orengo, C. A. Protein Eng 2000, 13, 153.
6. Dietmann, S.; Holm, L. Nat Struct Biol 2001, 8, 953.
7. Irving, J. A.; Whisstock, J. C.; Lesk, A. M. Proteins 2001, 42, 378.
8. McLachlan, A. D. Acta Crystallogr, Sect A 1972, 28, 656.
9. Kabsch, W. Acta Crystallogr, Sect A 1976, 32, 922.
10. Lathrop, R. H. Protein Eng 1994, 7, 1059.
11. Orengo, C. A.; Sillitoe, I.; Reeves, G.; Pearl, F. M. G. J Struct Biol 2001, 134, 145.
12. Taylor, W. R.; May, A. C. W.; Brown, N. P.; Aszodi, A. Rep Prog Phys 2001, 64, 517.
13. Eidhammer, I.; Jonassen, I.; Taylor, W. R. J Comput Biol 2000, 7, 685.
14. Lathrop, R. H. J Comput Biol 1999, 6, 405.
15. Orengo, C. A.; Taylor, W. R. J Theor Biol 1990, 147, 517.
16. Feng, Z. K.; Sippl, M. J. Fold Des 1996, 1, 123.
17. Godzik, A. Protein Sci 1996, 5, 1325.
18. Jaroszewski, L.; Rychlewski, L.; Godzik, A. Protein Sci 2000, 9, 1487.
19. Yang, A. S.; Honig, B. J Mol Biol 2000, 301, 679.
20. Carbó, R., Ed. In Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches; Kluwer Academic Publishers: Amsterdam, 1995.
21. Carbó-Dorca, R.; Mezey, P. G., Eds. In Advances in Molecular Similarity, volume 1; JAI PRESS Inc: Greenwich, CT, 1996.
22. Mezey, P. G. Shape in Chemistry: An Introduction to Molecular Shape and Topology; VCH: New York, 1993; pp 21–48.
23. Hohenberg, P.; Kohn, W. Phys Rev B 1964, 136, 864.
24. Parr, R. G.; Yang, W. Density-Functional Theory of Atoms and Molecules; Oxford University Press: New York, 1989; pp 51–53.
25. Mezey, P. G. Int Rev Phys Chem 1997, 16, 361.
26. Mezey, P. G. Int J Quantum Chem 1997, 63, 39.
27. Mezey, P. G.; Fukui, K.; Arimoto, S.; Taylor, K. Int J Quantum Chem 1998, 66, 99.
28. Mezey, P. G. J Chem Inf Comp Sci 1999, 39, 224.
29. Mezey, P. G. J Mol Mod 2000, 6, 150.
30. Constans, P.; Amat, L.; Carbó-Dorca, R. J Comput Chem 1997, 18, 826.
31. Woolley, R. G. J Am Chem Soc 1978, 100, 1073.
32. Carbó, R.; Leyda, L.; Arnau, M. Int J Quantum Chem 1980, 17, 1185.
33. Constans, P.; Carbó, R. J Chem Inf Comput Sci 1995, 35, 1046.
34. Constans, P.; Amat, L.; Fradera, X.; Carbó, R. In Advances in Molecular Similarity, Vol. 1; Carbó-Dorca, R.; Mezey, P. G., Eds.; JAI PRESS Inc: Greenwich, CT, 1996.
35. Vainshtein, B. K.; Fridkin, V. M.; Indenbom, V. L. Structure of Crystals, volume 2; Springer-Verlag: Berlin-Heidelberg, 1995; pp 453–456.
36. Coppens, P. X-Ray Charge Densities and Chemical Bonding; Oxford University Press: New York, 1997; pp 34–37.
37. Coulson, C. A.; Thomas, M. W. Acta Crystallogr, Sect B 1971, 27, 1354.
38. Barnes, J.; Hut, P. Nature 1986, 324, 446.
39. Greengard, L. Science 1994, 265, 909.
40. Strain, M. C.; Scuseria, G. E.; Frisch, M. J. Science 1996, 271, 51.
41. Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids; Oxford University Press: New York, 1990; pp 146–152.
42. Greenstadt, J. Math Comput 1967, 21, 360.
43. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535.
44. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235.
45. Constans, P. Tables of Atomic Densities from H to Kr. http://www.molspaces.com/dl/data/atdens.pdf (2001).
46. Constans, P. LSIM: A Quantum Macromolecular Similarity Program, v1.0. http://www.molspaces.com/lsim (2001).
47. Hsiou, Y.; Ding, J.; Das, K.; Clark, A. D.; Hughes, S. H.; Arnold, E. Structure 1996, 4, 853.
48. Das, K.; Ding, J. P.; Hsiou, Y.; Clark, A. D.; Moereels, H.; Koymans, L.; Andries, K.; Pauwels, R.; Janssen, P. A. J.; Boyer, P. L.; Clark, P.; Smith R. H.; Smith, M. B. K.; Michejda, C. J.; Hughes, S. H.; Arnold, E. J Mol Biol 1996, 264, 1085.
49. Scouloudi, H.; Baker, E. N. J Mol Biol 1978, 126, 637.
50. Kachalova, G. S.; Popov, A. N.; Bartunik, H. D. Science 1999, 284, 473.
51. Royer, W. E.; Heard, K. S.; Harrington, D. J.; Chiancone, E. J Mol Biol 1995, 253, 168.
52. Boel, E.; Brady, L.; Brzozowski, A. M.; Derewenda, Z.; Dodson, G. G.; Jensen, V. J.; Petersen, S. B.; Swift, H.; Thim, L.; Woldike, H. F. Biochemistry 1990, 29, 6244.
53. Brzozowski, A. M.; Davies, G. J. Biochemistry 1997, 36, 10837.
54. Reitzer, R.; Gruber, K.; Jogl, G.; Wagner, U. G.; Bothe, H.; Buckel, W.; Kratky, C. Struct Fold Des 1999, 7, 891.
55. Blake, J. D.; Cohen, F. E. J Mol Biol 2001, 307, 721.
56. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graphics 1996, 14, 33.
57. Merritt, E. A.; Bacon, D. J. Methods Enzymol 1997, 277, 505.