Bioinformatics, Vol. 14, No. 2, pp 196–205, 1998 (reformatted)

Estimation and filtering of potential protein–protein docking positions

Friedrich Ackermann* Grit Herrmann Stefan Posch
Gerhard Sagerer

Bielefeld University
Technical Faculty
P.O. Box 100 131
D–33501 Bielefeld
Germany.
email: {friedric|grit|posch|sagerer}@techfak.uni-bielefeld.de

Keywords: Protein docking, molecular recognition, augmented molecular surfaces, complementarity of molecular surfaces, vector valued cross correlation

July 3, 1998

Abbreviations used: AA, amino acid; RMSD, root mean square deviation; PDB, Brookhaven Protein Data Bank; FFT, fast Fourier transformation; 3d, three–dimensional; MAPC, method of aligned principal components; CFAC, check for aligned centroids

Abstract

Motivation: Software systems predicting automatically whether and how two proteins may interact are highly desirable, both, for understanding biological processes and for rational design of new proteins. As a part of a future complete solution to this problem a bundle of programs is presented designed (1) to estimate initial docking positions for a given pair of docking candidates, (2) to adjust them, and (3) to filter them, thus preparing more detailed computations of free energies.

Results: The system is evaluated on a test set of 51 cocrystallized complexes aiming at redocking the subunits. It works completely automatically and the evaluation is performed using one single set of parameters for all complexes in the test set. The number of solutions is fixed to 50 positions with a median CPU–time of 26min. For 30 complexes these contain a near–correct solution with RMSD $\leq$ 5.0Å, which is ranked first in five cases. For all complexes the best solution is scored on rank 16 as the worst case, and has a median RMSD of 4.3Å. Alternatively to this initial estimation of docking positions a global sampling of rotations was tested. Whereas this yields top–ranked solutions with RMSD $\leq$ 3.0Å for all 51 complexes the median CPU–time increases to 11h. This shows, that this blind sampling is not feasible for most applications.

Availability: The system and its components are available on request from the authors.

Contact: E-mail addresses: friedric@techfak.uni-bielefeld or posch@techfak.uni-bielefeld.de.

Introduction

Protein–protein interactions are one of the fundamental processes in living organisms, examples including interactions between receptors and peptide– or protein–hormones, enzymes and substrates or inhibitors, antibodies and antigens. Many biologically active proteins occur as multimeric associations, stabilised by protein–protein interactions between their individual subunits. In natural environment proteins associate spontaneously without external assistance if the free energy of the system of associated proteins and solvent has a lower free energy than the system of individually solvated proteins and solvent. Therefore it should be possible to decide theoretically whether two proteins will associate and if so in which relative position. Up to now, most algorithms devoted to solve this so called "protein docking problem", share four common principles (Cherfils and

*To whom correspondence should be addressed. Now with Behring Diagnostics GmbH, Instrumentation, Am Kronberger Hang 3, D–65824 Schwalbach/Ts., Germany.

Janin 1993):

1. They treat the proteins as rigid bodies, thus reducing the problems degrees of freedom to six, although it is known, that proteins docking with each other may adopt their 3d shapes ("induced fit"),
2. they rely on the detection of some relevant subsets of surface points potentially constituting the docking site,
3. they match these relevant surface points to obtain potential docking positions,
4. and they use scoring schemes based mainly on the geometrical complementarity of matched molecular surfaces instead of explicitly evaluating free energies.

Systems falling into this class of algorithms include DOCK (Kuntz et al. 1982; Shoichet and Kuntz 1991), the system of Connolly (1986) detecting critically shaped points, the geometric hashing systems of Norel et al. (1994) and Fischer et al. (1995), a system computing the critical points in an octree–representation of protein surfaces (Lenhof 1995), or most recently a system matching possible donors and acceptors in its initial step (Meyer et al. 1996).

The main idea behind the attempt to detect and to match relevant subsets of surface points is to reduce the size of the search space consisting of three parameters for the rotation and three parameters for the translation. Alternatively, either a complete sampling of all rotations and translations has to be performed or heuristic optimisation procedures like simulated annealing have to be applied (e.g. by Cherfils et al. 1991).

Additionally, docking systems differ with respect to the scoring scheme they use. Some employ only coarse criteria like maximising buried surface areas in the absence of steric hindrance (Katchalski-Katzir et al. 1992), others enhance the scores of good solutions distinguishing between favourable and unfavourable interactions at the contact site of the docked complex (Jiang and Kim 1991; Vakser and Aflalo 1994). As a final step an explicit evaluation of energy contributions can be performed as done by Jackson and Sternberg (1995) using initially estimated docking coordinates computed with the system of Shoichet and Kuntz (1991).

If tested on cocrystallized complexes docking systems exhibit a trade–off between the accuracy of the final result and the necessary CPU–time. Some systems compute potential solutions with RMSD well below 1.0Å, e.g. the systems of Totrov and Abagyan (1994) and Meyer et al. (1996). Their drawback are CPU–times of up to 30h for one complex (Totrov and Abagyan 1994). Another approach recovers solutions with RMSD up to 2.3Å decreasing CPU–times to 80min, on average (Lenhof 1995). Obviously another trade–off exists between the number of potential solutions and the minimum RMSD: Norel et al. (1994) end up with $2.1 \cdot 10^5$ solutions among them at least one exists with 1.9Å RMSD (on average) whereas Ackermann, Herrmann, Posch, and Sagerer (1995) compute only 382 solutions increasing RMSD to 6.4Å.

Some authors also tackle the far more difficult problem of docking individual uncomplexed structures (e.g. Shoichet and Kuntz (1991), Katchalski-Katzir et al. (1992), and others). Due to the phenomenon of "induced fit" difficulties result for all rigid body docking schemes, including ours. In the last paragraph possible solutions to the problem of induced fit within the framework of our system are outlined.

System overview

Fig. 1

The overall strategy of the docking system presented in this article conforms to the principles stated above. Its flow chart is shown in Figure 1. The main steps are

1. the computation of augmented molecular surfaces,
2. their subsequent segmentation into characteristically shaped regions,
3. the pairing of these regions leading to initial estimates of potential docking positions,
4. and the final adjustment and scoring of docking positions using a vector valued cross correlation.

The segmentation technique computes characteristically shaped regions on the surfaces of individual proteins (convexities, concavities). It is an extension of the method of Lee and Rose (1985) and has been reported in detail elsewhere (Meier et al. 1995): First the convex hull of the point set is computed, using a program by Barber et al. (1993). Now, the intrinsic features of the convex hull are exploited to detect convexities and to identify deep concave regions. The main ideas are:

1. A convexity is an area that protrudes from the surrounding surface with the tip lying directly on the facets of the convex hull.
2. A concavity is covered by large facets of the surface of the convex hull.

Region segmentation techniques adopted from image processing are used to implement the algorithm. Typically, approximately 10 convexities and 5 concavities are detected. Among them at least one overlaps approximately 60% with the true contact site of the cocrystallized complex. Therefore, the next step pairs a convexity and concavity from the surfaces of both proteins. To reflect saddle type interactions, we also allow pairings of two concavities. For each pair of regions potential docking positions are estimated and these positions are scored subsequently. In the following section the details of position estimation and scoring procedures are discussed.

Algorithmic details

The method of aligned principal components (MAPC)

Given a pair of regions R_1 and R_2 from two proteins, initial docking positions have to be estimated under the assumption that these two regions constitute the docking site of a complex. Without loss of generality we always fix the first protein and search for a possible translation vector $\vec{t}$ and a rotation matrix $\mathbf{R}$ to be applied to the second protein. This is done as follows: Each region $\nu = 1, 2$ is considered to be a set of 3d vectors $\vec{x}_i^\nu$. As the initial estimate of the translation vector $\vec{t}$ the difference between the centroids of both regions is used:

$$\vec{t} = \frac{1}{N_1} \sum_{i=1}^{N_1} \vec{x}_i^1 - \frac{1}{N_2} \sum_{i=1}^{N_2} \vec{x}_i^2 \tag{1}$$

Now, let $\vec{y}_i$ be the coordinates of surface points relative to the appropriate centroids

$$\vec{y}_i^\nu = \vec{x}_i^\nu - \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \vec{x}_i^\nu, \ \nu = 1, 2. \tag{2}$$

Then, the covariance matrix of the set of 3d surface points of a region is defined as

$$\mathbf{M}(R_\nu) = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \vec{y}_i^\nu (\vec{y}_i^\nu)^T, \ \nu = 1, 2. \tag{3}$$

Since a covariance matrix is symmetric it has three orthogonal eigenvectors which are called principal components of the region. The principal component $\vec{v}_{max}^\nu$ belonging to the largest eigenvalue of the covariance matrix is directed parallel to the maximal variance of the point set. Orthogonal to this, the principal component $\vec{v}_{min}^\nu$ belonging to the smallest eigenvalue of the covariance matrix is directed along the smallest variance of the region. The method of aligned principal components (MAPC) rotates the vectors $\vec{v}_{max}^2$ and $\vec{v}_{min}^2$ about a rotation axis $\vec{k}$ by an angle θ to align them with $\vec{v}_{max}^1$ and $\vec{v}_{min}^1$, respectively. $\vec{k}$ is determined enforcing the condition, that the projections of the rotated vectors onto $\vec{k}$ are invariant (cf. Figure 2), i.e.

$$\begin{aligned} \vec{k} \cdot \vec{v}_{max}^2 &= \vec{k} \cdot \vec{v}_{max}^1 \\ \vec{k} \cdot \vec{v}_{min}^2 &= \vec{k} \cdot \vec{v}_{min}^1. \end{aligned} \tag{4}$$

Thus, $\vec{k}$ and θ can be computed, and from this a rotation matrix $\mathbf{R}$ is obtained. Fig. 2

There remains an inherent ambiguity in the signs (+ or −) of principal components. Therefore, the MAPC leads to *four* initially estimated docking positions for each pair of regions corresponding to the four possible combinations of signs of principal components.

In the next subsection efficient scoring schemes are presented, well discriminating between the most correct and wrong solutions among these initially estimated docking positions.

Scoring initially estimated docking positions

Since each initially estimated docking position is derived from an unique pair of regions, the following three types of scoring functions are defined:

1. *Pair–independent scoring functions* attempt to classify each individual region, whether it possibly can constitute a docking site of the protein.
2. *Comparing scoring functions* compare the attributes of both regions of a potential pair and try to decide, whether they are complementary.
3. *Position–dependent scoring functions* evaluate whether a potential docking position, estimated on the basis of a certain pair of regions, is geometrically possible.

Initially estimated docking positions have to pass all three types of filters to be considered further by the vector valued cross correlation (see below). Pair–independent and comparing scoring functions are based on simple chemical and geometrical attributes. These are the sizes of regions and their mean hydrophobicities, computed using the empirical scale of Wolfenden et al. (1981). All thresholds applied are listed in Tables 1 – 3. They were determined empirically on the set of 51 cocrystallized complexes listed in the result section, where position–dependent scoring functions in addition depend on the position estimation technique applied. Segmented concavities have larger extent. All areas are within the ranges of known contact sites (Janin and Chothia 1990). The minimum sum of hydrophobicities of pairs of convexities with concavities is less than the same threshold for pairs of concavities with concavities, thus reflecting the polar interactions of protease–inhibitor complexes.

The position–dependent scoring functions usually show the best discriminative power. Among them the test for steric clash is quite obvious and frequently used also by other docking systems. As others do (Shoichet and Kuntz 1991; Lenhof 1995) for reasons of efficiency we compute the steric clash in discrete, grid–type data structures. Since even in rather coarse grids the test for steric clash is still a time consuming procedure (in our implementation needing approximately 300msec for each initially estimated docking position) beforehand the following model–based scoring function is evaluated, called the "check for aligned centroids (CFAC)". As indicated in Figure 3, a complementary pair of regions together with a correct position estimation leads to centroids of regions and proteins lying all four approximately on a straight line. In contrast, a wrong estimate leads to a geometrical constellation where the four points do not fall onto a straight line. Computing appropriate vector scalar products the necessary condition for a potential docking position to be a valid constellation can easily be checked.

Tab. 1

Tab. 2

Vector valued cross correlation

Tab. 3

Fig. 3

To adjust initially estimated docking positions a vector valued cross correlation is computed within dynamically determined subspaces. This technique combines for the first time the ideas

- to use FFT to compute the correlation (Katchalski-Katzir et al. 1992)
- with the enhancement of scores of good solutions by weighting surface complementarities by hydrophobicities (Vakser and Aflalo 1994) and partial charges
- together with the method of Meng et al. (1992) to save CPU–time by sampling protein volumes only in dynamically determined subspaces. These subspaces were called "clipping boxes" (Ackermann, Herrmann, Kummert, Posch, Sagerer, and Schomburg 1995). They surround the area in 3d space, where the proteins come in contact given the initially estimated docking position.

Fig. 4

As indicated in Figure 4 the aim of the cross correlation is to maximise the buried surface area without producing steric hindrance. To this end the cross correlation of the following three pairs of functions is computed. All functions are defined on a 3d grid with 32^3 or 64^3 grid points, where the size is chosen to obtain a sampling rate below 2Å. Sampling the proteins into the grids each grid point (i,j,k) is labelled with the appropriate function values: The geometry is reflected by the "volume–surface–functions" as they are used already by Katchalski-Katzir et al. (1992) and also by Meyer et al. (1996):

$$P_1(i,j,k) = \begin{cases} 1, & (i,j,k) \text{ surface protein 1} \\ 1, & (i,j,k) \text{ interior protein 1} \end{cases}$$

$$P_2(i,j,k) = \begin{cases} 1, & (i,j,k) \text{ surface protein 2} \\ -\rho, & (i,j,k) \text{ interior protein 2} \end{cases} \quad (5)$$

The cross correlation

$$P_1 \bullet P_2(i,j,k)$$
$$= \frac{1}{N^3} \sum_{i'j'k'} P_1(i',j',k') \cdot P_2(i+i', j+j', k+k') \quad (6)$$

will be maximal for geometrically perfect complementary surfaces, whereas steric hindrance results in negative penalty terms $1 \cdot (-\rho)$. Contacts of surface points with hydrophobicities exceeding the arithmetic mean $\bar{h} = -5.0$kcal/mol of the applied hydrophobicity scale (Wolfenden et al. 1981) are rewarded if the pair of "hydrophobicity surface functions" is cross correlated, where $h_\nu, \nu = 1,2$, are the hydrophobicities of the residues of protein ν:

$$H_1(i,j,k) = \begin{cases} \bar{h} + h_1, & (i,j,k) \;\; \begin{array}{l}\text{surface protein} \\ \text{1 and } h_1 \geq \bar{h}\end{array} \\ 0, & \text{else} \end{cases}$$

$$H_2(i,j,k) = \begin{cases} \bar{h} + h_2, & (i,j,k) \;\; \begin{array}{l}\text{surface protein} \\ \text{2 and } h_2 \geq \bar{h}\end{array} \\ 0, & \text{else} \end{cases} \quad (7)$$

Finally, the pointwise complementarity of opposite partial charges is reflected by cross correlating the pair of "surface charge functions" where $q_\nu, \nu = 1,2$, are the partial charges of the atoms of protein ν:

$$Q_1(i,j,k) = \begin{cases} q_1, & (i,j,k) \text{ surface protein 1} \\ 0, & (i,j,k) \text{ interior protein 1} \end{cases}$$

$$Q_2(i,j,k) = \begin{cases} -q_2, & (i,j,k) \text{ surface protein 2} \\ 0, & (i,j,k) \text{ interior protein 2} \end{cases} \quad (8)$$

The necessary partial charges are taken from Weiner et al. (1984).

To obtain a final scoring, the three contributing parts have to be weighted adequately. This is done with a convex combination

$$\begin{aligned} C(i,j,k) &= (1-\alpha)\cdot(1-\beta)\cdot(P_1 \bullet P_2)(i,j,k) \\ &+ \alpha\cdot(1-\beta)\cdot(H_1 \bullet H_2)(i,j,k) \\ &+ \beta\cdot(Q_1 \bullet Q_2)(i,j,k). \end{aligned} \quad (9)$$

The two parameters $\alpha \in [0, 1]$ and $\beta \in [0, 1]$ allow for an optimal tuning of the signal. Note, that the special case $\alpha = \beta = 0$ results in the original geometric cross correlation of Katchalski-Katzir et al. (1992). The initial estimation of docking positions was designed to avoid the time consuming blind sampling of rotations using the cross correlation. To compare our implementation with the original work of Katchalski-Katzir et al. (1992) we nevertheless tested a global search of rotational parameters. In this case, a sampling of the whole proteins with 128^3 grid points was used.

Implementation and results

The complete system consists of four programs, all written in C++. Two of them are usually used offline, computing the augmented surfaces and segmenting these surfaces. They are discussed in detail elsewhere (Ackermann 1997). As indicated in Figure 1 the third program pairs the segmented regions. It applies the scoring functions, predicts potential docking positions for each valid pair of regions, and finally sorts them according to the CFAC. The sorted list of potential docking positions is pruned, where the number of retained potential solutions is a parameter to be specified by the user.

In Table 4 all 51 cocrystallized complexes included in the experiments are listed. As requested in Norel et al. (1995) all tests are carried out "with a single set of parameters, without any further tuning whatsoever"! For each complex 2442 potential solutions are hypothesized (median) the most correct one having 8.5Å RMSD to the experimentally determined position (median). This solution is placed on rank 20 (median). Therefore as a cut–off for all 51 complexes the 25 highest scoring docking positions are retained as initial estimates.

For each complex these 25 solutions are fed into the program computing the vector–valued cross correlation. It tries to adjust each of them, optimizing the score $C(i, j, k)$ of the vector valued cross correlation (see Equation 9). For each initial docking position two adjusted positions with maximum score $C(i, j, k)$ are retained, thus leading to a fixed number of $25 \cdot 2 = 50$ potential docking positions for each complex. These final 50 solutions are ranked according to $C(i, j, k)$. In Table 4 final ranks and RMSD–values of the best solution for each of the 51 complexes are listed. Among the 50 solutions the most correct one has 4.3Å RMSD (median). This best solution is placed on rank 5 (median), worst case is rank 16 (complex 4XIA). In 30 out of 51 complexes at least one near–correct solution with RMSD $\leq$ 5.0Å is computed. Five top–ranked solutions exhibit RMSD $\leq$ 5.0Å (complexes 1MVPA–1MVPB, 2PTCE–2PTCI, 2TSCA–2TSCI, 3TPIZ–3TPII, and 5HVPA–5HVPB). The worst position estimations are those of complexes 2MCG, 3RUB and 6ADH (RMSD 18.6Å, 17.8Å and 14.4Å).

During the development of the system the weights of the convex combination of contributions to the cross correlation were varied systematically. As shown in Figure 5 the optimal result is only achieved, if both the hydrophobicity component $(H_1 \bullet H_2)(i, j, k)$ and the surface charge component $(Q_1 \bullet Q_2)(i, j, k)$ are taken into account. The optimal set of weight parameters is $\alpha_{opt} = 0.2$ and $\beta_{opt} = 0.5$. The median total CPU–times is 25.7min on a DEC 3000 AXP 300 with a maximum of 89.3min (complex 4XIA). Fig. 5

A linear correlation exists between the results of the initial position estimation and the final vector valued cross correlation (columns 6 and 9 of Table 4): The linear correlation coefficient is $c = 0.49$ and it is significant with an error probability of less than 0.1%.

The 51 complexes can roughly be split in two main types or classes: 29 of them containing two nearly equivalent chains are called the dimeric class. Among them typical examples are ubiquitin (1AARA: 76 AAs, 1AARB: 76 AAs), oxidoreductase (6ADHA: 374 AAs, 6ADHB: 374 AAs), immunoglobulin (2FBJH: 220 AAs, 2FBJL: 213 AAs), and uteroglobin (2UTGA: 70 AAs, 2UTGB: 70 AAs). Members of this class are printed boldface in Table 4. All other complexes consist of two clearly distinct chains, e.g. proteinase–inhibitor complexes (2SNIE: 275 AAs, 2SNII: 83 AAs) or complexes of an antibody with an antigen (1FDLH+1FDHLL: 432 AAs, 1FDLY: 129 AAs). It is obvious that the system behaves better on this rest class: For those 22 complexes a median of 2.7Å RMSD is achieved, where the median for the 29 complexes of the dimeric class is 5.1Å. This is mainly due to saddle type interactions and to complexes where the true contact site, i.e. the buried surface area in the cocrystallized complex, can be decomposed in two spatially disconnected components. This geometric situation is mainly observed at the interface of the light and heavy chains of antibodies (4FAB, 1FAIH, 1FDLH docked against 1FDLL etc.). Tab. 4

As stated above more precise docking positions can be obtained with global sampling of the complete rotational parameter space. For appropriate parameter sets, this results in a top–ranked solution with RMSD $\leq$ 3.0Å for all 51 complexes. Examples of RMSD and total computation time include 1CHO (0.56Å, 11.4h), 1MCP (1.50Å, 12.4h), 2PTC (2.56Å, 11.7h), 4FAB (1.78Å, 10.6h), and 6ADH (0.84Å, 13.0h). The docked subunits are as given in Table 4. For this global sampling $\alpha = \beta = 0$ may be chosen for optimal results yielding purely geometrical correlation of protein surfaces.

Discussion

The results reported in the previous section show the docking system to compare well to other systems on cocrystallized complexes. For example, in Norel et al. (1994) the best solution for each of the 16 considered complexes has RMSD below 5Å, with median computation time of 45min. However, this best solution is ranked 160 (median) and is scored on rank $1.4 \cdot 10^5$ in the worst case. On the other hand, (Meyer et al. 1996) report a median RMSD of 0.8Å for 45 complexes, which is placed on rank 1 for 41 complexes. The computation time for the initial estimate of potential docking position is 20min, where no timing requirements for the subsequent correlation are given. As a third example, (Lenhof 1995) redocks 52 cocrystallized complexes where the best solution is top–ranked with RMSD $\leq$ 5.0Å for 41 complexes. The median RMSD for the best solution is 1.9Å which is ranked first for 42 complexes and on rank 290 at the worst case, where one of the 52 complexes could not be docked at all. Mean computation time is 80min with 1163min as the worst case. These are the best results out of three different experiments with varying parameter sets.

Our system recovers original relative positions of two proteins with reasonable accuracy on a test set that has been systematically obtained by joining the test sets of three independent previous publications. The accuracy on a subset of test complexes called "dimeric class of complexes" consisting of two nearly homologous chains is worse than the accuracy on the rest class. However, it should be noted, that this rest class contains the test complexes most important for a possible application of the system (enzyme–inhibitor complexes, antibody–antigen complexes). We feel, that the typical member of the "dimeric class" — despite its use in all three publications (e.g. 2MHB) — is rather an "academic docking problem" in comparison.

A main feature of the presented system distinguishing it from previous approaches is its strictly bounded complexity: Sizes of solution space are restricted through all stages of the computation and the most correct solutions are ranked in the top of the list of estimated solutions. As a consequence a fixed number of hypothetical solutions is obtained and the CPU–times of the system lower the CPU–times of most previous approaches, both on average and in worst case. This is also due to the incorporation of spatially restricted grid sampling techniques in our system. Since we observe that lattice constants in 3d grid may not exceed a typical atom radius to make the cross correlation technique working, this method of dynamically determined subspaces is superior to the attempt to reduce the complexity of the cross correlation by the use of low resolution grids (Vakser 1995).

Estimation of positions and their scoring are the two main tasks of a docking system (Cherfils and Janin 1993). As in other systems these tasks can be identified with two stages of the computation, where the first one generates a set of possible solutions which are subsequently scored. We can clearly demonstrate, that scoring functions exist, sorting efficiently and reliably the predicted possible solutions — provided the most correct hypothetical solution has *low enough* RMSD to the correct position. Indeed, if the initial position estimation by the MAPC is replaced by a global sampling of relative orientations the cross correlation technique leads to practically perfect results with RMSD of top–ranked solutions $\leq$ 3.0Å for all 51 complexes. This has already be shown in the original paper of Katchalski-Katzir et al. (1992), it is reproduced impressively by Meyer et al. (1996) and we also can use our system in this way. One even does not need the vector valued cross correlation, the geometrical component alone will do. The only but crucial drawback of this *brute force* approach is its complexity: Already a step size of only 10° in each Eulerian angle produces approximately 23,000 nonequivalent orientations exceeding the number of initially estimated positions by the MAPC by a factor 10.

We see the rigid body docking as the "first step in addressing the more general (flexible) problem" (Norel et al. 1995). In this paper, neither the docking of flexible proteins nor the problem to dock individually crystallized subunits with a rigid body algorithm are addressed. Nevertheless, some of the most characteristic features of our system should fit well into a flexible docking scheme. First, the segmentation is stable on families of structurally similar proteins in the following sense (Biermann et al. 1996): As an example take the pancreatic trypsin inhibitor 2PTCI. In the database of families of structurally similar proteins (FSSP Ver. 01-May-1997, Holm et al. 1992) this protein, which is part of a complex, is member of the family indexed 1bpi. The representative protein 1BPI of this family is an individually crystallized inhibitor. If the regions segmented on the surfaces of 2PTCI and 1BPI are mapped onto each other using the transformation parameters from the FSSP, on the surface from 1BPI a region exists, nearly identical to the region exhibiting the largest overlap with the true contact site of the cocrystallized complex 2PTC. Therefore the starting point of our algorithm — the segmentation of characteristically shaped regions and their subsequent pairing — would lead to nearly identical results, both, for the cocrystallized inhibitor and for its individually crystallized relative. Second, the clipping box computed to restrict the complexity of the search space during the computation of the vector valued cross correlation could also serve to restrict the number of additional degrees of freedom, that have to be considered for a flexible docking scheme.

Interpreting our comparatively successful docking of trypsins and chymotrypsins with trypsin inhibitors (complexes 1CHOE–1CHOI, 1TGSZ–1TGSI, 2PTCE–2PTCI, 2TGPZ–2TGPI, 3TPIZ–3TPII, and 4TPIZ–4TPII), known to be difficult due to the presence of "false positives" (Shoichet and Kuntz 1991; Cherfils and Janin 1993), again it should be kept in mind, that we started with the complexed structures. Beside this, in 4 out of these 6 complexes the top–ranked solution is a "false positive" one (with RMSD > 5.0Å) in our docking algorithm, too.

The hydrophobic component and the surface charge component of the vector valued cross correlation mainly play the role to distinguish near correct (but still geometrically not completely perfect complementary) solutions from essentially wrong ones. Whereas on one hand it is trivial that additional information should lead to improved results, our observation proves, that an early incorporation of simple chemical filters can efficiently restrict the search space, thus facilitating more sophisticated energy evaluations. This is in perfect agreement with the results of Meyer et al. (1996), also indicating that chemical filters very successfully can be applied *before* tests for geometrical complementarity are performed.

As pointed out in the previous section, all results presented in this paper are obtained with one single and unified set of parameters for all complexes in the test set. On one hand this demonstrates the robustness of the system and its algorithms. This corresponds to the "pattern recognition point of view" to the docking problem: The rigid docking problem is formally equivalent to common machine vision and 3d object recognition problems (Norel et al. 1994), and on a given test set the recognition accuracy of any pattern recognition system can trivially be enhanced just by increasing the number of free parameters.

On the other hand, from a biological point of view, interfaces between proteins of different types of complexes *are* different both in geometrical types of interaction and in chemical features (Janin and Chothia 1990). One therefore could think of an extended version of the system presented so far, where a parameter set is chosen automatically depending on the types of proteins to be docked. It is expected, that an individual optimisation of the relative weights α and β in the convex combination of contributions to the vector valued cross correlation (Equation 9) leads to different settings α_{opt} and β_{opt} on different protein pairs thus further improving the overall results. Ackermann, Herrmann, Posch, and Sagerer (1995) present an early approach to this, and our future work will be directed this way.

Acknowledgements

Our work was funded by the German Ministry for Education, Science and Technology by grant no. 01 IB 307 C. We thank M. Meyer and D. Schomburg for many valuable hints and discussions.

References

Ackermann, F. (1997). *Prädiktion und Bewertung von Protein–Protein–Dockingpositionen*. Lage: Jacobs.

Ackermann, F., G. Herrmann, F. Kummert, S. Posch, G. Sagerer, and D. Schomburg (1995, Jul.). Protein docking combining symbolic descriptions of molecular surfaces and grid–based scoring functions. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak (Eds.), *Proceedings Third International Conference on Intelligent Systems for Molecular Biology*, pp. 3–11. Menlo Park: AAAI Press.

Ackermann, F., G. Herrmann, S. Posch, and G. Sagerer (1995, Oct.). 3D–Segmentierungstechniken und vektorwertige Bewertungsfunktionen für symbolisches Protein–Protein–Docking. In D. Schomburg and U. Lessel (Eds.), *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, pp. 105–124. Weinheim: VCH.

Barber, C. B., D. P. Dobkin, and H. Huhdanpaa (1993). The quickhull algorithm for convex hull. TR GCG53, Geometry Center, Minneapolis.

Biermann, R., F. Ackermann, and S. Posch (1996, Sept.). Robustness of geometrical shape descriptors for molecular surfaces. In *Proceedings German Conference on Bioinformatics (GCB'96)*, Leipzig, pp. 171–173. Universität Leipzig.

Cherfils, J., S. Duquerroy, and J. Janin (1991). Protein–protein–recognition analyzed by docking simulation. *PROTEINS: Structure, Function, and Genetics 11*, 271–280.

Cherfils, J. and J. Janin (1993). Protein docking algorithms: simulating molecular recognition. *Current Opinion in Structural Biology 3*, 265–269.

Connolly, M. L. (1986). Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers 25*, 1229–1247.

Fischer, D., S. L. Lin, H. Wolfson, and R. Nussinov (1995). A geometry–based suite of molecular docking processes. *Journal of Molecular Biology 248*, 459–477.

Holm, L., C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend (1992). A database of protein structure families with common folding motifs. *Protein Science 1*, 1691.

Jackson, R. M. and M. J. E. Sternberg (1995). A continuum model for protein–protein interactions: Application to the docking problem. *Journal of Molecular Biology 250*, 258–275.

Janin, J. and C. Chothia (1990). The structure of protein–protein recognition sites. *Journ. Biol. Chem. 265*, 16027.

Jiang, F. and S.-H. Kim (1991). "soft docking": Matching of molecular surface cubes. *Journal of Molecular Biology 219*, 79–102.

Katchalski-Katzir, E., I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA 89*, 2195–2199.

Kuntz, I. D., J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin (1982). A geometric approach to macromolecule–ligand interactions. *Journal of Molecular Biology 161*, 269.

Lee, R. H. and G. D. Rose (1985). Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers 24*, 1613–1627.

Lenhof, H.-P. (1995). An Algorithm for the Protein Docking Problem. Research Report MPI-I-95-1-023, Max–Planck–Institut für Informatik, Saarbrücken.

Meier, R., F. Ackermann, G. Herrmann, S. Posch, and G. Sagerer (1995, Oct.). Segmentation of molecular surfaces based on their convex hull. In *Proceedings International Conference on Image Processing*, Washington, D.C, pp. 552–555. IEEE Computer Society Press.

Meng, E. C., B. K. Shoichet, and I. D. Kuntz (1992). Automated docking with grid–based energy evaluation. *Journal of Computational Chemistry 13*, 505–524.

Meyer, M., P. Wilson, and D. Schomburg (1996). Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *Journal of Molecular Biology 264*, 199–210.

Norel, R., D. Fischer, H. Wolfson, and R. Nussinov (1994). Molecular surface recognition by a computer vision–based technique. *Protein Engineering 7*, 39–46.

Norel, R., S. Lin, H. Wolfson, and R. Nussinov (1994). Shape complementarity at protein–protein interfaces. *Biopolymers 34*, 933–940.

Norel, R., S. Lin, H. Wolfson, and R. Nussinov (1995). Molecular surface complementarity at protein–protein interfaces: The critical role played by surface normals at well placed, sparse, points in docking. *Journal of Molecular Biology 252*, 263–273.

Shoichet, B. K. and I. D. Kuntz (1991). Protein docking and complementarity. *Journal of Molecular Biology 221*, 327–346.

Totrov, M. and R. Abagyan (1994). Detailed *ab initio* prediction of lysozyme–antibody complex with 1.6å accuracy. *Nature Structural Biology 1*, 259–263.

Vakser, I. A. (1995). Protein docking for low–resolution structures. *Protein Engineering 8*, 371–377.

Vakser, I. A. and C. Aflalo (1994). Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *PROTEINS: Structure, Function, and Genetics 20*, 320–329.

Weiner, S. J., P. A. Kollmann, D. Case, U. Chandra Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society 106*, 765–784.

Wolfenden, R., L. Anderson, P. M. Cullis, and C. Southgate (1981). Affinities of amino acid side chains for solvent water. *Biochemistry 20*, 849.

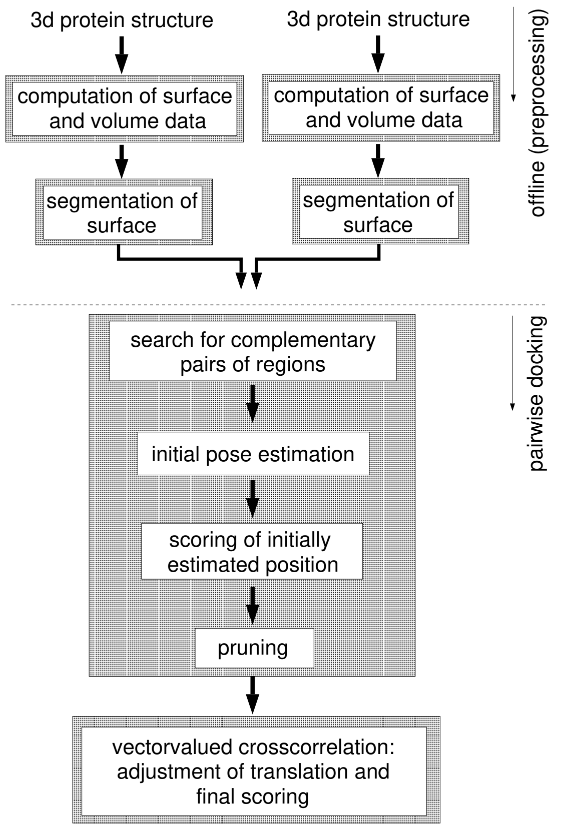

Figure 1: Flow chart of the system. Offline steps performed for individual proteins and their surfaces are followed by pairing of surface regions, position estimation and scoring procedures. The algorithm terminates with the vector valued cross correlation, adjusting the initially estimated positions and finally scoring them.

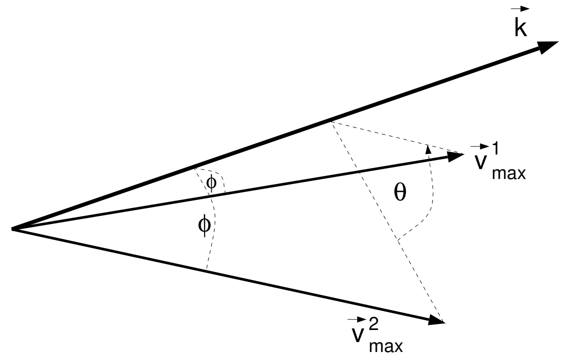

Figure 2: Estimating an axis of rotation $\vec{k}$ and an angle of rotation θ by the MAPC: The aim is to rotate $\vec{v}^2_{max}$ (cf. text for notation) in such a way, that it becomes aligned with $\vec{v}^1_{max}$ and to align in the same moment $\vec{v}^2_{min}$ with $\vec{v}^1_{min}$. $\vec{k}$ and θ can be determined by the requirement, that the projections of respective vectors along $\vec{k}$ (i.e. the angles ϕ) remain invariant under the rotation. Only the maximal principal components ($\vec{v}^2_{max}, \vec{v}^1_{max}$) are depicted for reasons of clarity.

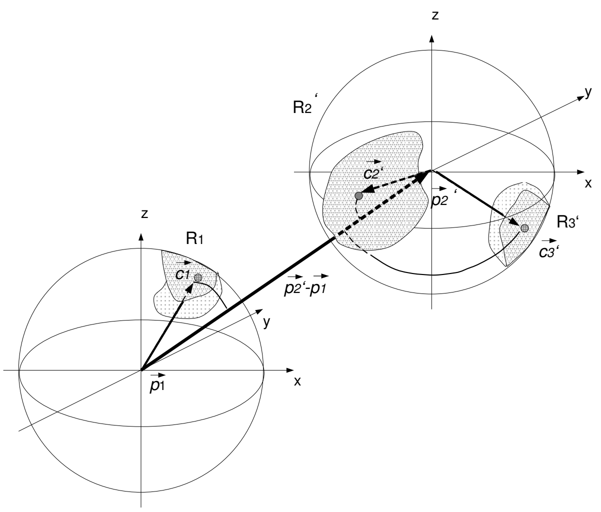

Figure 3: In the case of a well predicted docking position a relative constellation of centroids of regions and proteins results as indicated by the region R'_2 (primed coordinates to be taken after having applied the estimated transformation): the vectors $\vec{c}_1 - \vec{p}_1$ and $\vec{p}'_2 - \vec{c}'_2$ approximately are parallel to the vector $\vec{p}'_2 - \vec{p}_1$. The constellation for the region R'_3 shows an example for a possible erroneously estimated docking position.

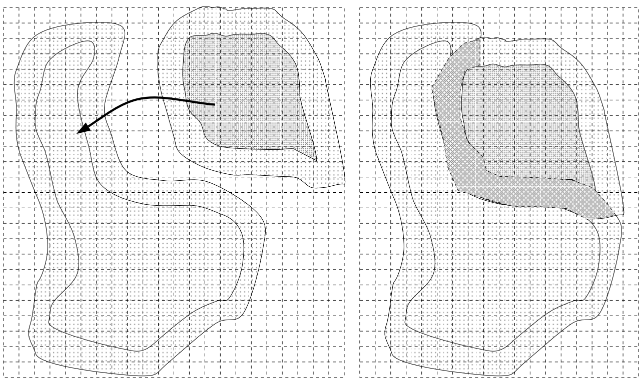

Figure 4: The principal idea in docking proteins via cross correlation techniques is indicated in this sketch: The two proteins are sampled in a 3d grid. Thus, the values at the grid points define a pair of functions, e.g. the volume–surface–functions of Equation 5. Now a translation is searched maximising the buried surface area but avoiding steric hindrance. This is done by cross correlating the pair of functions and searching for the peak of the cross correlation. If the hydrophobicity surface functions of Equation 7 or the surface charge functions of Equation 8 are used, the cross correlation reflects the pointwise complementarity of the corresponding chemical features.

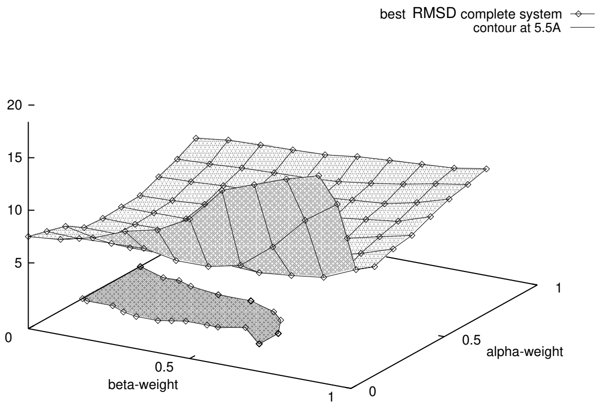

Figure 5: While developing the system, the relative weights of the correlations of hydrophobicity surface functions and surface charge functions where systematically varied. Relative weights $\alpha = \beta = 0.0$ lead to pure geometrical cross correlation functions. Each value $\alpha > 0.0$ or $\beta > 0.0$ moves the system towards a stronger relative contribution of the two chemical attributes. Since the penalty for steric clash is only reflected in the geometrical component it is clear that the optimal combination of parameters will be a set of values $\alpha < 1.0$ and $\beta < 1.0$. This figure shows as a surface the median RMSD of the most correct solution depending on α and β. This surface is completely continuous. In the $\alpha - \beta$–plane the area is indicated where RMSD of the most correct solution is less than 5.5Å. The optimum is within this stripe at $\alpha_{opt} = 0.2$ and $\beta_{opt} = 0.5$.

Attribute	concavity	convexity
Area [Å^2]	[19, 7188]	[11, 725]
Mean hydrophobicity	[1, 20]	[2, 23]

Table 1: Pair–independent scoring functions aim to classify whether a certain segmented region possibly constitutes a docking site. To do so, the attributes of the region have to fall in the intervals. Concavities and convexities are treated differently, since geometrical and chemical conditions differ between these two classes. Areas are measured in Å^2, the empirical hydrophobicity scale is dimensionless.

Attribute	concavity — convexity	concavity — concavity
Difference of areas [Å^2]	[26, 4283]	[13, 3230]
Sum of mean hydrophobicities	[14, 39]	[18, 37]

Table 2: Comparing scoring functions classify pairs of segmented regions yet without estimation of a relative position. A pair of regions passes the filters, if the value computed for the respective attribute falls into the interval, otherwise the pair is discarded. Pairs of convexities with concavities are considered to possibly constitute valid pairings as well as pairs of two concavities to reflect saddle–type interactions. Since their geometrical and chemical features differ, two different sets of intervals are determined.

Attribute	concavity — convexity	concavity — concavity
CFAC–cosine	> 0.1	> 0.1
Steric clash	$<$ 319 grid points	$<$ 354 grid points

Table 3: Position–dependent scoring functions for initially estimated docking positions rely on the thresholds listed above. A relative docking position has to be computed before they can be applied. As with the other scoring functions, the values of the attributes have to fall into the respective intervals, for a potential docking position to be classified as valid. Otherwise it is discarded. CFAC–cosine is the mean of normed scalar products of the relative region centroid vectors with the vector between the protein centroids (cf. Figure 3). Theoretically possible values range from –1.0 to 1.0. The steric clash is computed by counting grid points occupied by atoms of both proteins, given the estimated docking position. Each complex is sampled in a 3d regular grid with lattice constant 1.5Å.

1	2	3	4	5	6	7	8	9	10	11
PDB chains				initial position estimation			final scoring			total
Id	#AA	Id	#AA	rank	RMSD	time	rank	RMSD	time	time
1AARA	76	**1AARB**	76	14	5.6	243	5	6.3	728	971
1CHOE	245	1CHOI	56	123	5.2	291	3	2.9	790	1081
1CSEE	274	1CSEI	71	11	4.7	406	7	4.3	684	1090
1FAIH	221	**1FAIL**	214	15	15.7	206	2	3.2	2299	2505
1FDLH +1FDLL	432	1FDLY	129	13	13.0	939	12	5.2	1307	2246
1FDLH	218	**1FDLL**	214	4	10.2	204	2	5.1	2264	2468
1LYMA	129	**1LYMB**	129	127	12.7	256	4	11.6	798	1054
1MCPH	222	**1MCPL**	220	22	15.6	214	14	4.2	1810	2024
1MEEA	275	1MEEI	64	25	5.0	396	9	1.7	741	1137
1MVPA	124	**1MVPB**	124	50	9.3	449	1	4.6	1579	2028
1P01A	198	1P01P	4	5	6.5	21	9	6.7	980	1001
1P02A	198	1P02P	4	68	7.7	44	12	7.3	1015	1059
1PP2L	122	**1PP2R**	122	20	11.6	123	15	0.9	837	960
1TECE	279	1TECI	70	3	7.9	289	8	4.1	850	1139
1TGSZ	229	1TGSI	56	16	5.3	293	3	3.1	675	968
1TIMA	247	1TIMB	247	74	11.9	1109	5	3.0	1818	2927
2CCYA	128	**2CCYB**	128	130	11.8	195	13	6.4	866	1061
2FB4H	229	**2FB4L**	216	46	8.7	582	13	8.3	1875	2457
2FBJH	220	**2FBJL**	213	11	11.8	333	7	8.4	1558	1891
2KAIA +2KAIB	232	2KAII	58	68	7.1	243	9	3.8	599	842
2LTNA +2LTNB	233	**2LTNC** +2LTND	233	12	6.5	812	8	4.6	1477	2289
2MCG1	216	**2MCG2**	216	157	14.9	615	4	18.6	2405	3020
2MCPH	222	**2MCPL**	220	3	8.4	563	2	1.7	2522	3085
2MHBB	146	**2MHBA**	141	114	11.6	269	8	5.0	1281	1550
2PABA	127	**2PABB**	127	2	10.8	502	3	5.2	1037	1539
2PTCE	223	2PTCI	58	196	9.8	198	1	1.8	689	887
2RSPA	124	**2RSPB**	124	2	6.5	775	4	3.9	1703	2478
2SECE	274	2SECI	71	50	4.7	383	3	1.5	625	1008
2SNIE	275	2SNII	83	26	6.0	388	5	1.3	832	1220
2TGPZ	229	2TGPI	58	2	8.2	196	4	1.1	610	806
2TSCA	264	**2TSCB**	264	9	8.9	1719	1	0.8	2081	3800
2UTGA	70	**2UTGB**	70	95	6.8	96	5	13.2	613	709
3APRE	325	3APRI	8	19	2.8	39	7	5.8	664	703
3GPDG	334	**3GPDR**	334	11	6.6	1873	10	9.4	2583	4456
3HFMH +3HFML	419	3HFMY	129	15	14.3	918	5	2.7	1415	2333
3HFMH	215	**3HFML**	214	4	13.1	290	3	5.1	2900	3190
3HLAA	270	3HLAB	99	20	5.0	165	13	1.3	2049	2214
3RUBL	477	3RUBS	123	90	26.1	216	13	17.8	1836	2052

Table 4: Table of final results, part I. See explanation below part II.

1	2	3	4	5	6	7	8	9	10	11
PDB chains				initial position estimation			final scoring			total
Id	#AA	Id	#AA	rank	RMSD	time	rank	RMSD	time	time
3SGBE	185	3SGBI	56	42	8.1	92	13	5.6	505	597
3TIMA	250	**3TIMB**	250	4	13.4	1300	3	3.2	1148	2448
3TPIZ	229	3TPII	58	28	6.8	90	2	2.2	605	695
4CPA_	307	4CPAI	38	8	8.7	385	5	2.0	923	1308
4CTSA	437	**4CTSB**	437	39	10.4	443	1	8.6	3369	3812
4FABL	219	**4FABH**	216	66	14.8	502	13	9.2	2375	2877
4HVPA	99	**4HVPB**	99	129	5.5	502	10	3.3	621	1123
4SGBE	185	4SGBI	51	25	6.3	177	13	2.7	629	806
4TPIZ	229	4TPII	58	20	5.2	201	3	1.9	680	881
4XIAA	393	**4XIAB**	393	53	7.7	2356	16	7.9	3003	5359
5CSCA	433	**5CSCB**	433	5	14.0	940	3	2.1	2899	3839
5HVPA	99	**5HVPB**	99	1	6.3	167	1	4.9	748	915
6ADHA	374	**6ADHB**	374	36	16.6	2069	10	14.4	2804	4873
Medians (dimeric class)				20	10.8	502	5	5.1	1810	2457
Medians (rest class)				20	6.5	216	7	2.7	689	1008
Medians (total)				20	8.4	293	5	4.3	1037	1539

Table 4: Table of final results, part II. The first four columns list the 51 combinations of PDB–chains used in this test. Chains are concatenated, if the concatenation is the biological active unit (e.g. 2KAIA+2KAIB) or to test different combinations. Columns 5 and 6 list the results of the initial position estimation, columns 8 and 9 the final results. Column 7 notes the CPU–times necessary for the initial position estimation and scoring, column 10 the CPU–time for the cross correlation. The last column lists the total CPU–time including position estimation, scoring, and cross correlation. Times are given in sec and measured on a DEC 3000 AXP 300. Columns "rank" note the rank of the solution with minimum RMSD. Columns "RMSD" note the corresponding root mean square deviation in Å. The number of final solutions is fixed: For each of the 25 initially estimated docking position the two higest scoring adjusted positions are taking thus resulting in a final number of 50 positions for each complex. Chain–id's printed boldface indicate members of the so called "dimeric class" of complexes consisting of two nearly homologous chains. This class contains 29 complexes. All 22 other complexes are collected in the rest class. In five cases the top–scored solution exhibits RMSD ≤ 5.0Å: 1MVPA–1MVPB, 2PTCE–2PTCI, 2TSCA–2TSCI, 3TPIZ–3TPII, and 5HVPA–5HVPB.