



Taylor & Francis
Taylor & Francis Group

Journal ... **Optimization
Methods and Software**

Article ID ... **GOMS220297**

TO: CORRESPONDING AUTHOR

**AUTHOR QUERIES - TO BE ANSWERED BY THE
AUTHOR**

Dear Author

Please address all the numbered queries on this page which are clearly identified on the proof for your convenience.

Thank you for your cooperation

	No queries.	

Production Editorial Department, Taylor & Francis
4 Park Square, Milton Park, Abingdon OX14 4RN

Telephone: +44 (0) 1235 828600
Facsimile: +44 (0) 1235 829000

Global optimization in protein docking using clustering, underestimation and semidefinite programming

ROUMMEL F. MARCIA*[†], JULIE C. MITCHELL[†] and STEPHEN J. WRIGHT[‡]

[†]Departments of Biochemistry and Mathematics, Mitchell Lab, 433 Babcock Drive,
UW-Madison, Madison, WI 53706, USA

[‡]Department of Computer Sciences, 120 West Dayton Street,
UW-Madison, Madison, WI 53706, USA

(Received)

The underestimation of data points by a convex quadratic function is a useful tool for approximating the location of the global minima of potential energy functions that arise in protein-ligand docking problems. Determining the parameters that define the underestimator can be formulated as a convex quadratically constrained quadratic program and solved efficiently using algorithms for semidefinite programming (SDP). In this paper, we formulate and solve the underestimation problem using SDP and present numerical results for active site prediction in protein docking.

Keywords: Protein docking; Global optimization; Convex underestimation; Semidefinite programming

AMS Subject Classification: 65K05, 65K10; 90C22, 90C30, 90C90

1. Introduction

Protein-ligand docking problems in computational biology can be formulated as global minimization problems in which the docked configuration of the two molecules corresponds to the global minimum of a potential energy function describing the molecular interaction. Typically, the energy landscape is funnel-shaped and highly nonlinear with many local minima, making the docking problem very difficult to solve. Convex global underestimators (CGU) [1–5] were developed to determine the location of the global minimum of such functions by iteratively underfitting a set of data points over a contracting domain by a sequence of strictly convex quadratic functions. The (unique) minimum of each convex quadratic underestimator is an approximation to the global minimum and is used to define the domain in the next iteration. Underestimation methods have been successfully implemented in protein structure determination [1–4] as well as in protein docking [6–10].

Determining the parameters p that define the underestimator $q(p; x)$ involves solving a nonlinear program that minimizes L^1 distance between the energy function $f(x^{(k)})$ and $q(p; x^{(k)})$,

*Corresponding author. Tel.: (608) 262-6844; Fax: (608) 262-3453; Email: marcia@math.wisc.edu

51 $k = 1, 2, \dots, m$, subject to $q(p; x^{(k)}) \leq f(x^{(k)})$ for each k ; see ref. (1). The L^1 distance is used
 52 primarily because it is robust with respect to outliers [11–13]. The quadratic function $q(p; x)$
 53 is required to underestimate $f(x)$ so that its minimizer is a plausible predictor of the global
 54 minimum of $f(x)$. To guarantee that $q(p; x)$ is convex, its Hessian matrix H with respect to
 55 x must be positive definite. In the original formulation of CGU [1], H was chosen to be a
 56 diagonal matrix with positive diagonal entries, which simplifies the underestimation problem
 57 to a linear program. The choice for the Hessian was generalized to dense positive definite
 58 matrices in ref. [5], where H was expressed in terms of its Cholesky factorization $H = LL^T$,
 59 with L a lower triangular matrix with positive diagonal entries. In this case, the nonlinear
 60 program becomes a quadratically constrained quadratic program whose global solution is not
 61 necessarily easy to find. A two-phase approach, proposed in ref. [5], finds an underestimator
 62 with a diagonally dominant Hessian H in the first phase, using H as the initial guess in the
 63 second phase, which solves for the Cholesky factorization LL^T . The globally optimal positive
 64 definite Hessian is almost always found by this approach. In this paper, we solve the underesti-
 65 mation problem using a single-phase approach, formulating it as a semidefinite programming
 66 (SDP) problem that can be solved efficiently.

67 SDP can be viewed as an extension of linear programming in which the unknowns include
 68 symmetric matrices as well as vectors and scalars, and nonnegativity constraints on the
 69 variables become positive semidefiniteness requirements on the symmetric matrix variables.
 70 More information on SDP can be found in ref. [14] and ref. [15]. It is natural to formulate
 71 (1) as an SDP in which the Hessian of the quadratic underestimator q appears as one of the
 72 unknowns. Paschalidis *et al.* [16], concurrently with the submission of this paper, described
 73 an underestimation procedure based on an SDP formulation, using a biased sampling proce-
 74 dure to select new starting points at each major iteration. The authors discuss the relationship
 75 of their algorithm to earlier versions of CGU [4], and obtain test results on three problems,
 76 including those used in examples 4 and 5 below. However, they fixed the three orientation
 77 variables for the ligand at their optimal values, and applied their procedure only to the three
 78 translational variables, and from a nearby starting point. In contrast, our results above did not
 79 assume any such knowledge of the global minimizer.

80 A second innovation presented in this paper is the use of clustering to determine separate
 81 regions of space to be searched for the global minimizer. Frequently, an initial scan of the
 82 parameter space reveals widely separated regions with low function values, which can be iden-
 83 tified by applying clustering procedures to the local minima found during the scan. A process
 84 of repeated convex underestimation, random point generation, and local optimization can be
 85 applied to each cluster separately, to find the ‘global’ minimizer for each cluster. The best such
 86 solution becomes our estimate of the overall global solution. (We note that clustering was not
 87 used in ref. [16].)

88 The paper is organized as follows. We describe the convex quadratic underestimator and
 89 the SDP formulation in section 2. Section 3 gives details on docking mesh evaluator (DoME),
 90 the algorithm for global minimization of energy functions arising in docking applications, and
 91 section 4 describes computational tests on five problems from the Protein Data Bank [17].
 92 We summarize our conclusions in section 5.

93 94 95 96 2. SDP formulation

97
 98 Given m data points $(x^{(1)}, f(x^{(1)}))$, $(x^{(2)}, f(x^{(2)}))$, \dots , $(x^{(m)}, f(x^{(m)}))$, where $x^{(k)} \in \mathfrak{R}^n$, $k =$
 99 $1, 2, \dots, m$, and $f: \mathfrak{R}^n \rightarrow \mathfrak{R}$, we compute a convex quadratic function that underestimates
 100 these data points by defining a quadratic function $q(c_0, c, H; x) = c_0 + c^T x + \frac{1}{2} x^T H x$, where

101 $c_0 \in \Re, c \in \Re^n$, and $H \in \Re^{n \times n}$, and solving the minimization problem

$$\begin{aligned}
 &102 \\
 &103 \quad \text{minimize}_{c_0, c, H} \quad \sum_{k=1}^m s_k \\
 &104 \\
 &105 \quad \text{subject to} \quad q(c_0, c, H, x^{(k)}) + s_k = f(x^{(k)}), \quad k = 1, 2, \dots, m \\
 &106 \\
 &107 \quad \quad \quad s_k \geq 0, \quad k = 1, 2, \dots, m \\
 &108 \quad \quad \quad H \text{ symmetric positive definite,} \\
 &109 \quad \quad \quad |c_0|, \|c\|, \text{ and } \|H\| \text{ bounded,} \\
 &110
 \end{aligned} \tag{1}$$

111 where $s_k \in \Re$ for $k = 1, 2, \dots, m$. The constraint that H is positive definite ensures that
 112 $q(c_0, c, H; x)$ is convex with respect to x . We impose the following explicit bounds on c_0
 113 and c :

$$114 \quad |c_0| \leq \beta_0, \quad \text{and} \quad \|c\|_\infty \leq \beta_c, \tag{2}$$

115 and discuss the bounding of $\|H\|$ further below.

116 We now formulate (1) as a semidefinite program. First, by introducing the slack variables
 117 $v_0^+, v_0^- \in \Re$ and $v^+, v^- \in \Re^n$, we can express (2) as

$$\begin{aligned}
 &118 \quad \quad \quad v_0^+ = \beta_0 - c_0 \geq 0 \\
 &119 \quad \quad \quad v_0^- = c_0 + \beta_0 \geq 0 \\
 &120 \quad \quad \quad v^+ = \beta_c e - c \geq 0 \\
 &121 \quad \quad \quad v^- = c + \beta_c e \geq 0, \\
 &122 \\
 &123 \\
 &124
 \end{aligned}$$

125 where $e \in \Re^n$ is the vector of ones. Note that

$$126 \quad c_0 = \frac{1}{2}(v_0^- - v_0^+), \quad \text{and} \quad c = \frac{1}{2}(v^- - v^+).$$

127 Using these formulae, we can eliminate the parameters c_0 and c , and the constraints (2) can
 128 be written as follows:

$$\begin{aligned}
 &129 \\
 &130 \\
 &131 \quad \quad \quad v_0^+ + v_0^- = 2\beta_0 \\
 &132 \quad \quad \quad v^+ + v^- = 2\beta_c e \\
 &133 \\
 &134 \quad \quad \quad v_0^+, v_0^-, v^+, v^- \geq 0.
 \end{aligned}$$

135 Next, we note that $x^T H x = X \cdot H$, where $X = x x^T$ and $A \cdot B = \sum_{i,j} A_{i,j} B_{i,j}$ is the standard
 136 inner product of symmetric matrices. To ensure that q has a (strictly) positive definite Hessian,
 137 we replace H by $H + \varepsilon I_n$, where $\varepsilon > 0$ is some parameter that is a lower bound to the smallest
 138 eigenvalue of the desired Hessian. The corresponding equality constraint becomes

$$139 \quad c_0 + c^T x^{(k)} + \frac{1}{2}(X^{(k)} \cdot H) + s_k = f(x^{(k)}) - \frac{\varepsilon}{2} \|x^{(k)}\|_2^2, \quad s_k \geq 0.$$

140 The size of H can be controlled by adding a regularization term $\lambda \cdot \text{trace}(H) = \lambda \cdot (I_n \cdot H)$ to
 141 the objective, where λ is a ‘tuning’ parameter that can be successively increased to force $\|H\|$
 142 to drop below a prescribed bound. In fact, λ can be interpreted as a Lagrange multiplier for
 143 a constraint $\text{trace}(H) \leq C$, for some bound C . (In our computational experience, $\|H\|$ does
 144 not grow too large even for $\lambda = 0$, but we include it in our formulation for completeness.)
 145 In summary, our SDP formulation of equation (1) is given by

$$146 \quad \text{minimize} \quad \sum_{k=1}^m s_k + \lambda (I_n \cdot H) \tag{SDP}$$

147
148
149
150

$$\begin{aligned}
& \text{subject to } \frac{1}{2}(v_0^- - v_0^+) + \frac{1}{2}(v^- - v^+)^T x^{(k)} + \frac{1}{2}(X^{(k)} \cdot H) + s_k \\
& \quad = f(x^{(k)}) - \frac{\varepsilon}{2} \|x^{(k)}\|_2^2 \text{ for all } k \\
& \quad v_0^+ + v_0^- = 2\beta_0 \\
& \quad v^+ + v^- = 2\beta_c e \\
& \quad v_0^+, v_0^-, v^+, v^- \geq 0 \\
& \quad s_k \geq 0, \quad k = 1, 2, \dots, m, \\
& \quad H \geq 0.
\end{aligned}$$

Although, we can apply transformations to the vector and scalar variables to obtain a problem with symmetric matrix variables only, such transformations are computationally inefficient and in any case unnecessary, as current SDP software can handle such variables explicitly.

3. The DoME algorithm

The DoME is software for predicting the active site of proteins upon binding with ions, DNA, ligands, and other macromolecules. The proteins are treated as rigid bodies, with one (receptor) fixed in space whereas the other (ligand) is allowed to move and rotate, leading to six degrees of freedom: three translational and three rotational. Previously, DoME described the molecular interaction by defining an energy model based only on solvent effects and van der Waals forces [7]. Specifically, the electrostatic interactions in DoME were described by a finite-element solution to the Poisson–Boltzmann equations (PBE) whereas the dipole moments and steric repulsion were computed using a standard Lennard–Jones 6–12 formula. Because the computed solution to the PBE was piecewise-linear, the potential energy function was not differentiable, and therefore, non-gradient-based optimization had to be used for local minimization.

In the new version of DoME, the potential energy function has been made differentiable by utilizing the analytical solution to the linearized PBE, known as the Yukawa potential. Whereas this model does not address dielectric effects to the same level of detail as the Poisson–Boltzmann model, it nonetheless incorporates the Debye–Huckel screening of charges in solvents. Two additional terms, representing hydrogen bonding and desolvation, have also been included in this version of DoME. Solvation plays an important role in protease-inhibitor complexes. The hydrogen bond energy term has a hydrogen-acceptor cutoff distance of 4.5 Å and a donor-hydrogen-acceptor (D-H-A) cutoff angle of 90°. It includes a dependence on the D-H-A angle since bond strength favors a linear alignment. The free energy associated with removing solvents at the active site is approximated by computing the change in the solvent accessible surface area [18]. (For further details on each energy term, see ref. [19].)

DoME searches for the global minimum by performing an exhaustive scanning preprocessing phase, followed by a sequence of major iterations, each consisting of scanning and underestimation. In the preprocessing phase, a scan is performed over the six degrees of freedom by holding the receptor in a fixed position and orientation, then placing the ligand at various distances from it (fixing three degrees of freedom) and oriented in various ways (fixing the other three). Typically, energies are evaluated at about 2 million sampling points during this initial phase, and the lowest 900 are used as starting points for local optimizations. The local minima thus found are underfitted with the convex quadratic function constructed in the manner described in section 2. The global minimizer of the underestimator is used to initialize

201 another local search and thus obtain another local minimizer, which we call x_{pred}^* . On each
202 subsequent major iterations, we construct a new, smaller search domain that encompasses x_{pred}^*
203 along with the lowest k_b local minima obtained from the other local searches. (A typical value
204 for k_b is three.) Random points are generated within this smaller domain, followed by local
205 minimization from each of these points, underestimation, and another local search form the
206 minimizer of the underestimator. The process is repeated until the predicted global minima
207 is close to the local minima with the lowest known function value and the domain size is
208 sufficiently small. This algorithm is described more fully in ref. [7].

209 It has been shown that the coupled use of scanning and optimizing is more effective in
210 determining points of low-energy values than by scanning or optimization alone [6]. However,
211 we find that the initial exhaustive scan sometimes produces points with low-energy values
212 in distinct regions in space. In this work, we enhanced DoME by clustering the low-energy
213 points found during the scan in the preprocessing phase on the basis of their (x, y, z) co-
214 ordinates. We then form a separate quadratic underestimator for each cluster and apply the
215 subsequent major iterations to each cluster separately. At the end of this process, we obtain
216 an approximate global minimizer for each cluster. For clustering, we used the k -Median
217 algorithm of Bradley *et al.* [20], which minimizes the total L^1 distance between points in a
218 chosen cluster. This algorithm is robust and is guaranteed to converge to a solution satisfying
219 the minimum principle necessary optimality condition for the problem [21]. No more than
220 four clusters were used, ensuring that each cluster contained sufficiently many points for
221 underestimation. This approach differs from the clustering algorithm for protein complexes
222 called ClusPro [22], which computes the root-mean-square deviation (RMSD) between ligand
223 residues at the interface of each candidate conformations and groups those structures within
224 a (default) cluster radius of 9 Å RMSD. The clusters are then ranked based on the number
225 of structures they contain. It is possible that such an approach would yield many clusters (as
226 many as 30) and would not guarantee the number of data points necessary to determine the
227 convex quadratic underestimator, so we elected not to use this approach.

230 4. Computational examples in protein docking

231
232 We applied the enhanced DoME code described in the previous section to five examples from
233 the Protein Data Bank [17]. For each of these examples, the bound configuration is known,
234 allowing us to compare the DoME results with the known global minimizer.

235
236 *Example 1: CheY-binding domain of CheA in complex with CheY (1A0O).* The 1A0O
237 complex consists of response regulator of bacterial chemotaxis, CheY, bound to the recogni-
238 tion domain from its cognate histidine kinase, CheA [23]. This specific recognition domain
239 minimizes the cross-talk in signal transmission mediated by Mg^{2+} -dependent phospho-relay
240 reactions between histidine auto-kinases and phosphoaccepting receiver domains.

241
242 *Example 2: V-1 Nef protein in complex with wild type Fyn SH3 domain (1AVZ).*
243 The antibody-antigen complex 1AVZ consists of a viral protein, HIV-1 Nef, and the host cell
244 signal transduction protein, Fyn tyrosine kinase SH3 domain [24]. The interaction between
245 these two proteins provides for long-term survival of infected T cells and for destruction of
246 non-infected T cells by inducing apoptosis.

247
248 *Example 3: Trypsin complex with Bowman-Birk inhibitor (1TAB).* The 1TAB complex
249 consists of the enzyme trypsin and BBI, the Bowman-Birk trypsin-inhibitor, which is a
250 polypeptide chain of 71 amino acids highly cross-linked by seven disulfide bridges [25].

Table 1. Results for 1A00.

Cluster	Iter	$E(x_{\min}^0)$	$E(x_{\min}^f)$	$\ \cdot\ _2$	$\ \cdot\ _{\theta}$
1	4	-29.652	-41.779	32.516	1.314
2	3	-58.535	-58.804	0.138	0.022
3	4	-42.855	-46.526	55.774	3.093
4	5	-41.645	-44.416	55.690	2.954
-	4	-58.535	-58.788	0.122	0.025

Table 2. Results for 1AVZ.

Cluster	Iter	$E(x_{\min}^0)$	$E(x_{\min}^f)$	$\ \cdot\ _2$	$\ \cdot\ _{\theta}$
1	3	-31.071	-45.214	24.565	2.075
2	4	-42.132	-45.280	24.855	1.836
3	4	-45.515	-53.381	6.353	2.987
4	4	-47.406	-76.947	0.276	0.079
-	5	-47.406	-60.195	5.556	2.893

Table 3. Results for 1TAB.

Cluster	Iter	$E(x_{\min}^0)$	$E(x_{\min}^f)$	$\ \cdot\ _2$	$\ \cdot\ _{\theta}$
1	5	-41.503	-44.450	36.134	2.767
2	3	-50.438	-59.761	0.092	0.081
3	3	-47.359	-47.587	30.139	0.567
4	3	-52.275	-53.503	36.360	1.104
-	5	-52.275	-53.524	36.392	1.108

Elevated levels of trypsin have been found in pancreatic tumors, and BBI, commonly found in soybeans, has been shown to suppress this type of tumor in various animals.

Example 4: Barnase–barstar complex (1BRS) Barnase is an extracellular ribonuclease found in bacillus amyloliquefaciens. The intracellular polypeptide inhibitor barstar disrupts its potentially lethal functions by sterically blocking its active site with a helix and adjacent loop segment [26].

Example 5: Trypsin-pancreatic enzyme inhibitor complex (2PTC). The 2PTC complex involves trypsin interacting with a bovine pancreatic enzyme inhibitor [27].

We solved SDP using the MATLAB software package SDPT3 (version 3.0) of Tütüncü *et al.* [28], which is an interior-point algorithm that uses a predictor-corrector primal-dual path-following method. In our numerical testing, the default HKM [29–31] direction was used. All runs were made on a single 2.20 GHz Pentium 4 processor Linux workstation with 896 MB of RAM from Dell Computers and 2.0 GHz Apple Power Mac G5 Cluster. The computational times for each cluster in each example ranged between 7.5 and 12.5 h on four processors, with 80% of the time spent on local optimization and 20% on random point generation and function evaluations. Constructing the underestimator at each iteration took not more than two seconds.

Results for the five examples appear in Tables 1–5. We denote the local minima with the lowest energy value in the initial scanning and in the final iteration by x_{\min}^0 and x_{\min}^f , respectively. For each of the four clusters, we list the number of major DoME iterations

Table 4. Results for 1BRS.

Cluster	Iter	$E(x_{\min}^0)$	$E(x_{\min}^f)$	$\ \cdot\ _2$	$\ \cdot\ _{\theta}$
1	5	-40.996	-47.611	44.068	1.193
2	4	-49.732	-71.465	0.001	0.110
3	6	-52.534	-52.550	43.042	2.666
4	4	-41.650	-42.118	33.400	3.011
-	4	-52.534	-52.553	43.040	2.666

Table 5. Results for 2PTC.

Cluster	Iter	$E(x_{\min}^0)$	$E(x_{\min}^f)$	$\ \cdot\ _2$	$\ \cdot\ _{\theta}$
1	4	-41.383	-61.791	0.008	0.133
2	5	-45.805	-53.242	39.491	2.292
3	4	-47.584	-48.696	46.499	2.632
4	4	-49.924	-53.255	30.916	1.243
-	6	-49.924	-55.015	37.828	0.624

(Iter), the energy values ($E(x_{\min}^0)$ and $E(x_{\min}^f)$) in kcal/mol, and the 2-norm $\|\cdot\|_2$ (Å) and angular distances $\|\cdot\|_{\theta}$ (radians) of the local minima x_{\min}^f to the known docked configuration. As a point of comparison, we list in the last row of each table, indicated by ‘-’ in the Cluster column, the results obtained by scanning and underfitting without clustering the data points in the initial iteration.

4.1 Analysis of results

In the five examples presented, DoME approximated the locations of the global minima by clustering the initial points followed by iterative underestimation. In the 1A0O complex (Example 1), a point near the global minimum was detected in DoME’s initial scanning, in cluster 2. Naturally, DoME identified this point correctly as the global minimum, after several iterations. It is interesting to note that DoME also finds the global solutions if clustering is not used. Due to the low-energy value of the minimum ($E(x_{\min}^0) = -58.525$ kcal/mol) in relation to the other local minima values found, the shape of the underestimator is such that its minimum nearly coincides with the global minimum. Thus, there was a large decrease in the size of the search domain after the initial scanning phase, and subsequent iterations were able to obtain nearby local minima with even lower energy values.

In example 2 (1AVZ), the global minimum was again found in the initial scanning but was removed from the set of local minimizers before performing the underestimation, to see if subsequent iterations could recover this point. Again, clustering and iterative underestimation was able to locate a very good approximation to the global minimum, even though the initial point x_{\min}^0 in cluster 4 had a much higher energy value than the global minimizer, which lies in this cluster. When clustering was omitted in this example, DoME found a minimizer that was located relatively near the global minimum (about 5.5 Å away) but was oriented incorrectly. It is the steric repulsion induced by this incorrect orientation that prevented the ligand molecule from achieving the correct translational coordinates for binding.

In examples 3, 4, and 5 (the 1TAB, 1BRS, and 2PTC complexes), widely separated regions of low energy are identified during the initial scan, and the use of clustering enabled DoME to identify the global minimizer correctly in each case. When clustering was not used, the single quadratic underfitting function had its minimizer between these low-energy regions, in

351 a region that was neither low in energy value nor near any of the minima. In each of these
352 cases, when a local search was performed from the minimizer of the convex underestimator,
353 the best of the four initial local minimizers from each of the clusters was identified as x_{pred}^* .
354 Subsequent DoME iterations in the single-cluster case failed to identify the known global
355 minimizers.

356 It is interesting to note that our ‘divide and conquer’ approach to optimization always
357 produced a near-native (globally minimizing) structure as the predicted global minimum of
358 one of the initial clusters. In two of five cases, the near-native configuration corresponded
359 was found in the cluster whose local minimizer had the lowest function value after the initial
360 scanning phase. However, in the other three cases (1TAB, 1BRS, 2PTC) the near-native
361 structure was obtained in a different cluster, one for which the local minimizer obtained
362 after the scanning phase had a higher function value. This fact demonstrates the value of the
363 multiple cluster CGU approach, because it allows for several domains to be searched, thereby
364 increasing the chance of producing a near-native solution.

367 5. Conclusion

368 We presented a method for computing a convex quadratic function that underestimates a set
369 of points for determining the global minimum of a function. We formulated the problem
370 as a semidefinite program, which generally can be solved efficiently in theory and practice.
371 We applied this approach in the context of protein docking and showed that a combination of
372 clustering and iterative underestimation effectively predicted near-native docking configura-
373 tions for several test cases. A more comprehensive survey of docking problems will determine
374 whether near-native conformations are always found as cluster global minima.

378 Acknowledgements

379 The authors would like to thank Susan Lindsey and Erick Butzlaff for their contributions to
380 DoME. This work was supported by the Department of Energy Grant DE-FG03-01ER25497
381 and the National Library of Medicine Training Grant 5T15LM007359-03.

384 References

- 385 [1] Phillips, A., Rosen, J. and Walke, V., 1996, Molecular structure determination by convex global underestimation
386 of local energy minima. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* Vol. 23,
387 (American Mathematical Society), pp. 181–198.
- 388 [2] Dill, K.A., Phillips, A.T. and Rosen, J.B., 1997, CGU: an algorithm for molecular structure prediction. In *Large-*
389 *Scale Optimization with Applications, Part III* (Minneapolis, MN, 1995), *IMA Mathematics and its Applications*,
390 Vol. 94 (New York, Springer) pp. 1–21.
- 391 [3] Rosen, J.B., Dill, K.A. and Phillips, A.T., 1997, Protein structure and energy landscape dependence on sequence
392 using a continuous energy function. *Journal of Computational Biology*, **4**, 227–239.
- 393 [4] Phillips, A.T., Rosen, J.B., Dill, K.A., 2001, Convex global underestimation for molecular structure prediction.
394 In: P.M. Pardalos et al. (Ed.) *From Local to Global Optimization* (Dordrecht, Netherlands: Kluwer Academic
395 Publishers), pp. 1–18.
- 396 [5] Rosen, J.B. and Marcia, R.F., 2004, Convex quadratic approximation. *Computational Optimization and*
397 *Applications*, **28**(2), 173–184.
- 398 [6] Mitchell, J.C., Rosen, J.B., Phillips, A.T. and Ten Eyck, L.F., 1999, Coupled optimization in protein docking.
399 In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ACM Press,
400 280–284.
- [7] Marcia, R.F., Mitchell, J.C. and Rosen, J.B., 2005, Iterative convex quadratic approximation for global
optimization in protein docking. *Computational Optimization and Applications*, **32**(3), 285–297.
- [8] Mangasarian, O.L., Rosen, J.B. and Thompson, M.E., 2005, Global minimization via piecewise-linear
underestimation. *Journal of Global Optimization*, **32**(1), 1–9.

- 401 [9] Mangasarian, O.L., Rosen, J.B. and Thompson, M.E., 2006, Convex kernel estimation of functions with multiple
402 local minima. *Computational Optimization and Applications*, **34**(1), 35–45.
- 403 [10] Mangasarian, O.L., Rosen, J.B. and Thompson, M.E., 2006, Nonconvex piecewise-quadratic underestimation
404 for global minimization. *Journal of Global Optimization*, **34**(4), 475–488.
- 405 [11] Barrodale, I., 1968, L1 approximation and the analysis of data. *Applied Statistics*, **17**, 51–57.
- 406 [12] Barrodale, I. and Roberts, F.D.K., 1973, An improved algorithm for discrete L1 linear approximation. *SIAM*
407 *Journal of Numerical Analysis*, **10**, 839–848.
- 408 [13] Rosen, J.B., Park, H., Glick, J. and Zhang, L., 2000, Accurate solution to overdetermined linear equations with
409 errors using L1 norm minimization. *Computational Optimization and Applications*, **17**(2–3), 329–341.
- 410 [14] Nesterov, Y. and Nemirovskii, A., 1994, Interior-point polynomial algorithms in convex programming,
411 volume 13 of SIAM Studies in Applied Mathematics (Philadelphia, PA: Society for Industrial and Applied
412 Mathematics (SIAM)).
- 413 [15] Vandenberghe, L. and Boyd, S., 1996, Semidefinite programming. *SIAM Review*, **38**(1), 49–95.
- 414 [16] Paschalidis, I.Ch., Shen, Y., Vajda, S. and Vakili, P., 2004, A semi-definite programming-based underestimation
415 method for global optimization in molecular docking. Proceedings of the 44th IEEE Conference on Decision
416 and Control and the European Control Conference 2005, 3675–3680.
- 417 [17] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr. Brice, M.D., Rogers, J.R., Kennard, O.,
418 Shimanouchi, O.T. and Tasumi, M., 1977, The Protein Data Bank: a computer-based archival file for
419 macromolecular structures. *Journal of Molecular Biology*, **112**, 535–542.
- 420 [18] Hasel, W., Hendrikson, T.F. and Still, W.C., 1988, A rapid approximation to the solvent accessible surface areas
421 of atoms. *Tetrahedron Computational Method*, **1**, 103–16.
- 422 [19] Schlick, T., 2002, *Molecular Modeling and Simulation* (New York: Springer-Verlag).
- 423 [20] Bradley, P.S., Mangasarian, O.L. and Street, W.N., 1997, Clustering via concave minimization. In: Michael C.
424 Mozer, Michael I. Jordan and Thomas Petsche (Eds), *Advances in Neural Information Processing Systems*,
425 Vol. 9 (The MIT Press), pp. 368.
- 426 [21] Bennett, K.P. and Mangasarian, O.L., 1993, Bilinear separation of two sets in n-space. *Computational*
427 *Optimization and Applications*, **2**(3), 207–227.
- 428 [22] Comeau, S.R., Gatchell, D.W., Vajda, S. and Camacho, C.J., 2004, ClusPro: an automated docking and
429 discrimination method for the prediction of protein complexes. *Bioinformatics*, **20**(1), 45–50.
- 430 [23] Welch, M., Chinardet, N., Mourey, L., Birck, C. and Samama, J.P., 1998, Structure of the CheYbinding domain
431 of histidine kinase CheA in complex with CheY. *Nature Structural Biology*, **5**, 25–29.
- 432 [24] Arold, S., Franken, P., Strub, M.P., Hoh, F., Benichou, S., Benarous, R. and Dumas, C., 1997, The crystal
433 structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered
434 T cell receptor signaling. *Structure*, **5**, 1361–1372.
- 435 [25] Tsunogae, Y., Tanaka, I., Yamane, T., Kikkawa, J., Ashida, T., Ishikawa, C., Watanabe, K., Nakamura, S. and
436 Takahashi, K., 1986, Structure of the trypsin-binding domain of Bowman-Birk type protease inhibitor and its
437 interaction with trypsin. *Journal of Biochemistry Tokyo*, **100**, 1637–1646.
- 438 [26] Buckle, A.M., Schreiber, G., Fersht, A.R., 1994, Protein-protein recognition: crystal structural analysis of a
439 barnase-barstar complex at 2.0-Å resolution. *Biochemistry*, **33**, 8878–8889.
- 440 [27] Marquart, M., Walter, J., Deisenhofer, J., Bode, W. and Huber, R., 1983, The geometry of the reactive site and
441 the peptide groups in trypsin, trypsinogen, and its complexes with inhibitors. *Acta Crystallography, Section B*,
442 **39**, 480–490.
- 443 [28] Tütüncü, R.H., Toh, K.C. and Todd, M.J., 2003, Solving semidefinite-quadratic-linear programs using SDPT3.
444 *Mathematical Programming serial B*, **95**(2), 189–217 (Computational semidefinite and second order cone
445 programming: the state of the art).
- 446 [29] Helmsberg, C., Rendl, F., Vanderbei, R.J. and Wolkowicz, H., 1996, An interior-point method for semidefinite
447 programming. *SIAM Journal on Optimization*, **6**(2), 342–361.
- 448 [30] Kojima, M., Shindoh, S. and Hara, S., 1997, Interior-point methods for the monotone semidefinite linear
449 complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, **7**(1), 86–125.
- 450 [31] Monteiro, R.D.C., 1997, Primal-dual path-following algorithms for semidefinite programming. *SIAM Journal*
451 *on Optimization*, **7**(3), 663–678.