Parameter Refinement for Molecular Docking

Jukka-Pekka Salo,* Ari Yliniemelä, and Jyrki Taskinen

Division of Pharmaceutical Chemistry, Department of Pharmacy, P.O. Box 56 (Viikinkaari 5),
FIN-00014 University of Helsinki, Finland

Received March 19, 1998

Finding the optimal parameter values for any computer program with adjustable parameters can be very time consuming. In this paper, we introduce the use of the Plackett−Burman and the central composite designs with the aid of the partial least squares method to tackle this problem. Using DOCK3.5 as a test case, we also show a four-step procedure for sequential docking utilizing two parameter sets, both effecting a different level of accuracy. The DOCK parameter values were refined for protein kinase C regulatory domain yielding an orientation at the global "energy" minimum, which is in very good agreement with the experimental protein kinase C regulatory domain−phorbol 13-acetate complex. The scheme is now being used for screening molecular databases to find putative protein kinase C inhibitors.

INTRODUCTION

DOCK3.5[1] is a suite of computer programs developed for discovering pharmacologically active molecules, which are complementary to the receptor, or enzyme, binding site,[2] by screening large molecular databases of three-dimensional structures. A prerequisite of using the program is that the 3D structure of the target protein (or the relevant subunit thereof) be known; homology based receptor or enzyme models may also be used.[3] During the docking, thousands of possible orientations for each ligand are systematically scanned by matching the distances between receptor target points to ligand atom−atom distances with a predefined tolerance,[4] and each orientation is evaluated with a simple and, subsequently, fast scoring function. The ligand orientation at the binding site is then refined by a simplex minimizer and rescored.[5]

DOCK is capable of identifying experimentally known ligand binding modes,[4−9] and it has been successfully applied for screening three-dimensional structural databases.[3,10−15] The fraction of compounds possessing good scorings has a remarkable enrichment of ligands with a desired pharmacological activity: DOCK clusters compounds that have affinity to the target protein with a typical hit rate of 2−20% for ligands possessing micromolar activity.[2] However, the binding mode suggested for an active molecule may not always be correct.[11,14] Kuntz' group[5] has validated DOCK by reproducing the experimental binding mode of several ligand−protein complexes. They have achieved orientations within 5 kcal/mol about the global "energy" minimum, but the optimal parameter set apparently varies from case to case. As DOCK is used for screening large databases, it is obviously impossible to validate the procedure by showing that the parameters are optimal for all ligand structures. When the binding mode of a high-affinity ligand is known, DOCK could be optimized—by rigorously reproducing the orientation for that ligand at the global "energy" minimum—to recognize its key features effecting the discovery of new structural leads possessing the correct features. For this approach, an effective refinement procedure for the parameters is needed.

DOCK is by no means the only computer program with a large set of adjustable parameters. With all of them, however, one fundamental issue persists: which path to follow in order to optimize the parameter values efficiently and reliably. Used for a long time, the straightforward approach is to optimize one factor at a time.[16] It is, however, rather inefficient and unreliable.[17−19] The same applies to the trial-and-error method, especially in a complex environment.[20] Also, either method obviously provides poor means to explain parameter interactions and are aimed toward ending up at the optimum proper rather than providing predictive power.[19,20] To reflect these considerations, the safest approach would be to explore the entire parameter space, which essentially means a grid of test points or a factorial design.[16,18,20] However, these designs soon become prohibitively large in size as the number of variables and tested levels increase.[16−18,20] *E.g.*, there are well over 20 optimizable parameters even for the smallest set of programs in DOCK3.5[1] as shown in Table 1. Thus, using a two-level factorial experiment to optimize their values would result in at least 2^{20} experiments! Therefore, a method is needed to identify the significant variables for a closer analysis.

The Plackett−Burman design[17] (PBD) is a fractional factorial-like screening method based on balanced incomplete blocks, which can be set up to yield a "saturated" design, *i.e.*, to investigate $n-1$ variables in n experiments. PBD also has the attractive feature of being capable of including both quantitative and qualitative parameters in the analysis, and both variable types can appear in the same problem.[21,22] One of the drawbacks of (fractional) factorials is that data could be gathered in a region of little interest (including too small/large a range[18]) if the workers have no prior knowledge of the relevant factor levels,[23] and due attention should be paid to choosing the range of each variable. This is a problem to be avoided by the use of the various simplex methods.[24] On the other hand, a longer time span may be required to

* To whom all correspondence should be addressed.

S0095-2338(98)00182-6 CCC: $15.00
Published on Web 08/19/1998

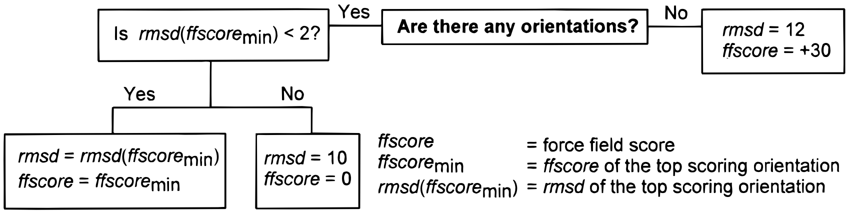

Is rmsd(ffscore_min) < 2?
Yes
Are there any orientations?
No
rmsd = 12
ffscore = +30
Yes
No
rmsd = rmsd(ffscore_min)
ffscore = ffscore_min
rmsd = 10
ffscore = 0
ffscore = force field score
ffscore_min = ffscore of the top scoring orientation
rmsd(ffscore_min) = rmsd of the top scoring orientation

Figure 1. The flow chart for evaluating the test runs during the optimization steps. Ligand orientations with an *rmsd* > 2 were considered very different from the orientation in the X-ray structure and were given arbitrary unfavorable response values to emphasize this fact.

go through the sequential optimization steps if time-consuming runs are involved. Furthermore, simplex methods are known to have been stuck at a local minimum.[20]

PBD was used in this work to identify the parameters with the highest significance, whose values would then be refined by analyzing the results of a central composite design[16] (CCD) with the aid of the partial least squares method[25] (PLS). We show that managing the results of a CCD using a number of pseudovariables in the PLS analysis yields very good values for the parameters of DOCK3.5[1] judged by the program's ability to find the apparent "energy" minimum for and to reproduce the crystal orientation of phorbol 13-acetate (PRB) on protein kinase C.[26] We further compare these results with those obtained by using a systematic search for *some* parameters as described by Gschwend *et al*.[5] and suggest a way to combine these methods into a four-step procedure, which is then used for screening large databases to enhance the reliability of the results by refining the idea of sequential docking published[12] previously.

METHODS

Crystal Structures. The crystal structures of protein kinase Cδ Cys2 domain complexed with phorbol 13-acetate[26] (1ptr), lactate dehydrogenase with NAD-lactate[27] (5ldh), dihydrofolate reductase with methotrexate[28] (3dfr), and catechol *O*-methyltransferase with 3,5-dinitrocatechol[29] (1vid) were taken from the Protein Data Bank.[30] Hydrogens were added in "favorable" orientations on the protein by the program *addprh*.[1] 1ptr served as the primary test case; 5ldh, 3dfr, and 1vid were used to test the parameter generality only.

Predocking Procedures. The Connolly surface[31] for the ligand binding site was calculated with the program *autoMS*.[4,31] The spheres in the binding pocket, whose centers serve as the receptor target points, were generated with the program *sphgen*.[4,32] Since the binding site was known, the size of the sphere cluster was reduced by removing all sphere centers farther than 3.0 Å from the crystal orientation of the ligand (with hydrogens added in standard orientations and Gasteiger−Marsili charges calculated in Sybyl 6.3[33]). Furthermore, all sphere centers closer than 2.3 Å (the minimum value tested for *polcon*, a parameter for the receptor polar atom close contacts of the program *chemgrid*[7]) to the heavy atoms of the protein were removed since *dock3.5*[5,6,8,34] tries to match only ligand heavy atoms on the sphere centers. The force field grid was calculated with the program *chemgrid*[7] to enclose the selected sphere centers with a 3 Å margin in each direction.

Parameter Refinement. The parameters *nodes_maximum* and *nodes_minimum*, which define the maximum and minimum numbers of ligand-atom/sphere-center pairs in generating a ligand orientation, were considered to be only one variable and were thus given equal values. The simplex minimization was used to refine the ligand orientations. *rmsd* values for the docked orientations (from the crystal orientation) were calculated by the program *dock3.5*.[5,6,8,34] The best force field score and the *rmsd* value associated with it were used as response unless otherwise stated; the exact procedure is described in Figure 1.

Two sequential refinement cycles were performed as depicted in Table 1; Plackett−Burman design was used to recognize the significant variables for the central composite design. The first cycle comprised finding a set of "decent" parameter values for the docking procedure itself (program *dock3.5*[5,6,8,34]); spread across the binding site, a cluster of 36 spheres from a preliminary test was used. After that, the optimization of the steps to be performed before the actual docking could be evaluated. In the second cycle, the only feasible response was the *rmsd* value of the best scoring orientation since the parameters of *chemgrid*[7] were being optimized. In the end, some systematic refinement of the parameters was performed to minimize the analysis time while maintaining accuracy.

Plackett−Burman Design.[17] Figure 2 shows the two-level PBD table for 16 experiments (up to 15 variables). For each experiment, one has a response r_i, *i.e.*, some measured value of significance, which can be used to calculate the main effects m_j of the variables

$$m_j = \frac{\sum r_i(+) - \sum r_i(-)}{n} \tag{1}$$

where $r_i(+)$ are the r_i's of the experiments with a "+" in the respective column, $r_i(-)$ are the r_i's of the experiments with a "−", and n is the number of experiments. The relative significance of each variable is given by $|m_j|$, and the sign of m_j reveals the direction of change in response when variable j moves from "−" to "+". In our case, negative values of m_j are desirable since smaller values of both the *rmsd* and the force field score represent a more favorable orientation of the ligand.

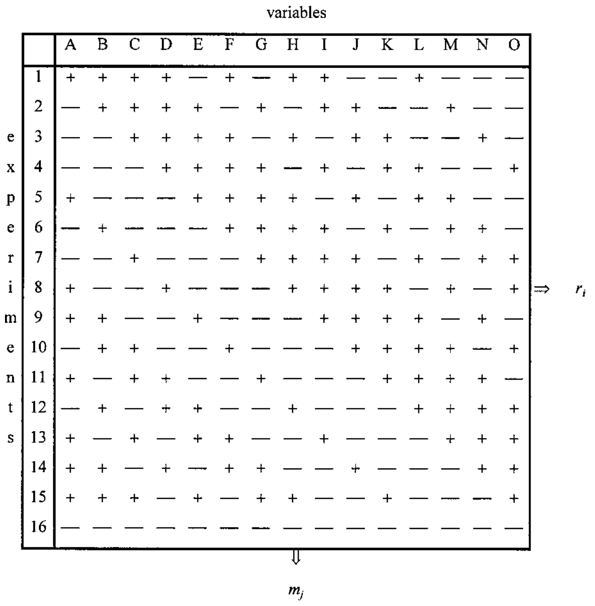
The number of variables investigated can be less than $n-1$. A number of variables can be assigned to be "dummies", which are used to evaluate the variance of error[17] or as an indication of interactions between variables. The significance of the main effects can be determined with the *t*-test, but this information is usually less important since further optimization is to be performed with the variables with the greatest $|m_j|$'s.[21]

Central Composite Design.[16] The principle of CCD is highlighted in Figure 3. It is made up of a star design ($2n+1$ experiments) and a two-level factorial design (n^2 experi-

Table 1. Optimizable Parameters in DOCK3.5[1] [a]

			PBD (−/+)			
program	variable	default value	cycle 1	cycle 2	refined value	fast value
autoMS[4,31]	*surface_density* (dots/Å^2)	3.0		**3.0**/9.0	3.0	3.0
	probe_radius (Å)	1.4		1.4/**1.0**	1.0	1.0
sphgen[4,32]	*dotlim*	0.0		**0.0**/0.2	0.0	0.0
	radmax (Å)	4.0−5.0		**4.0**/5.0	5.0	5.0
	radmin (Å)	*probe_radius*		*probe_radius*/***probe_radius*-0.3**	0.7	0.7
chemgrid[7]	*grddiv* (Å)	0.30		0.30/**0.15**	smallest possible value	
	estype	distance dependent		distance dependent		
	esfact	4		4.0/**4.5**	4.5	4.5
	cutoff (Å)	10.0		**10.0**/7.0	10.0	10.0
	pcon (Å)	2.3		**2.3**/2.0	2.3	2.3
	ccon (Å)	2.8		2.8/**2.5**	2.5	2.5
			Matching Parameters			
dock 3.5[5,6,8,34]	*distance_tolerance* (Å)	1.5	1.5/1.0		1.5	0.7
	nodes_maximum	4			4	4
	nodes_minimum	4	4/8		4	4
	ligand_binsize (Å)	1.0	1.0/0.5		0.5	0.4
	ligand_overlap (Å)	0.0	0.0/0.2**ligand_binsize*		0.25	0.3
	receptor_binsize (Å)	1.0	1.0/0.5		0.5	0.4
	receptor_overlap (Å)	0.0	0.0/0.2**receptor_binsize*		0.25	0.3
	bump_maximum	0	0/2		5	2
			Scoring Parameters			
dock 3.5[5,6,8,34]	*interpolate*	yes	yes/no		yes	yes
	minimize	no	yes		yes	yes
	simplex_iterations	500	25/100		500	500
	simplex_convergence	0.2	0.2/0.5		0.05	0.2
	simplex_restart	1.0	1.0/0.5		0.2	1.0
	simplex_initial_translation	1.0	1.0/2.0		1.0	1.0
	simplex_initial_rotation	0.5	0.5/1.0		0.5	0.5

[a] Additionally, there are four parameters for *distmap*[6,34] and eight parameters for *dock3.5*[5,6,8,34] not used in this work. Boldface values in PBD cycle 2 mark the settings in the best experiment.


variables

experiments		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	1	+	+	+	+	—	+	—	+	+	—	—	+	—	—	—
	2	—	+	+	+	+	—	+	—	+	+	—	—	+	—	—
e	3	—	—	+	+	+	+	—	+	—	+	+	—	—	+	—
x	4	—	—	—	+	+	+	+	—	+	—	+	+	—	—	+
p	5	+	—	—	—	+	+	+	+	—	+	—	+	+	—	—
e	6	—	+	—	—	—	+	+	+	+	—	+	—	+	+	—
r	7	—	—	+	—	—	—	+	+	+	+	—	+	—	+	+
i	8	+	—	—	+	—	—	—	+	+	+	+	—	+	—	+
m	9	+	+	—	—	+	—	—	—	+	+	+	+	—	+	—
e	10	—	+	+	—	—	+	—	—	—	+	+	+	+	—	+
n	11	+	—	+	+	—	—	+	—	—	—	+	+	+	+	—
t	12	—	+	—	+	+	—	—	+	—	—	—	+	+	+	+
s	13	+	—	+	—	+	+	—	—	+	—	—	—	+	+	+
	14	+	+	—	+	—	+	+	—	—	+	—	—	—	+	+
	15	+	+	+	—	+	—	+	+	—	—	+	—	—	—	+
	16	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Figure 2. The PBD test table for 16 experiments.[17] − notes the "lower" level and + the "higher" level of the variable. One need not attach any physical meaning to the expressions "lower" and "higher", and it is customary to use the usual or default values as the lower level.

ments). This design provides the analyst with information on five different levels for each variable.

Partial Least Squares. The PLS package in Sybyl 6.3[33] was used to derive a response function from the results of the CCD. For all the variables x_i in the CCD, x_i^2 through x_i^6 were calculated. Similarly, all two-factor, *e.g.*, x_1x_2, three-factor, and, where applicable, four-factor interactions were calculated. Additionally, all possible combinations of the form $x_1^2x_2$ and, where applicable, $x_1^2x_2x_3$ were calculated. The initial variable matrix comprised all these "pseudovariables" with the original variables; the target variable matrix consisted of the force field scores of the top scoring orientations. PLS validation was done by the leave-one-out method. Figure 4 shows the flow chart for our procedure. The results were compared to those with matrix algebra for the full second-order polynomial model and by a systematic search. The minima of the response functions were solved with Maple V.[35]

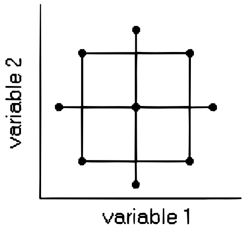


Figure 3. The principle of CCD for two variables, the dots represent experiments. $n^2 + 2n + 1$ experiments are needed to carry out the procedure for n variables.

Fast Parameters. The systematic approach described previously[5] to optimize the bin sizes (*ligand_binsize* = *receptor_binsize*), overlaps (*ligand_overlap* = *receptor_overlap*) and the matching distance tolerance (*distance_tolerance* = *binsize* + *overlap*) was used.

Docking Experiment. The same set of approximately 500 randomly selected molecules from the ACD-3D database[36] was docked using both our refined parameters and the fast parameters, and their rankings were recorded. Gasteiger−Marsili charges for the ligands were calculated in Sybyl 6.3.[33]

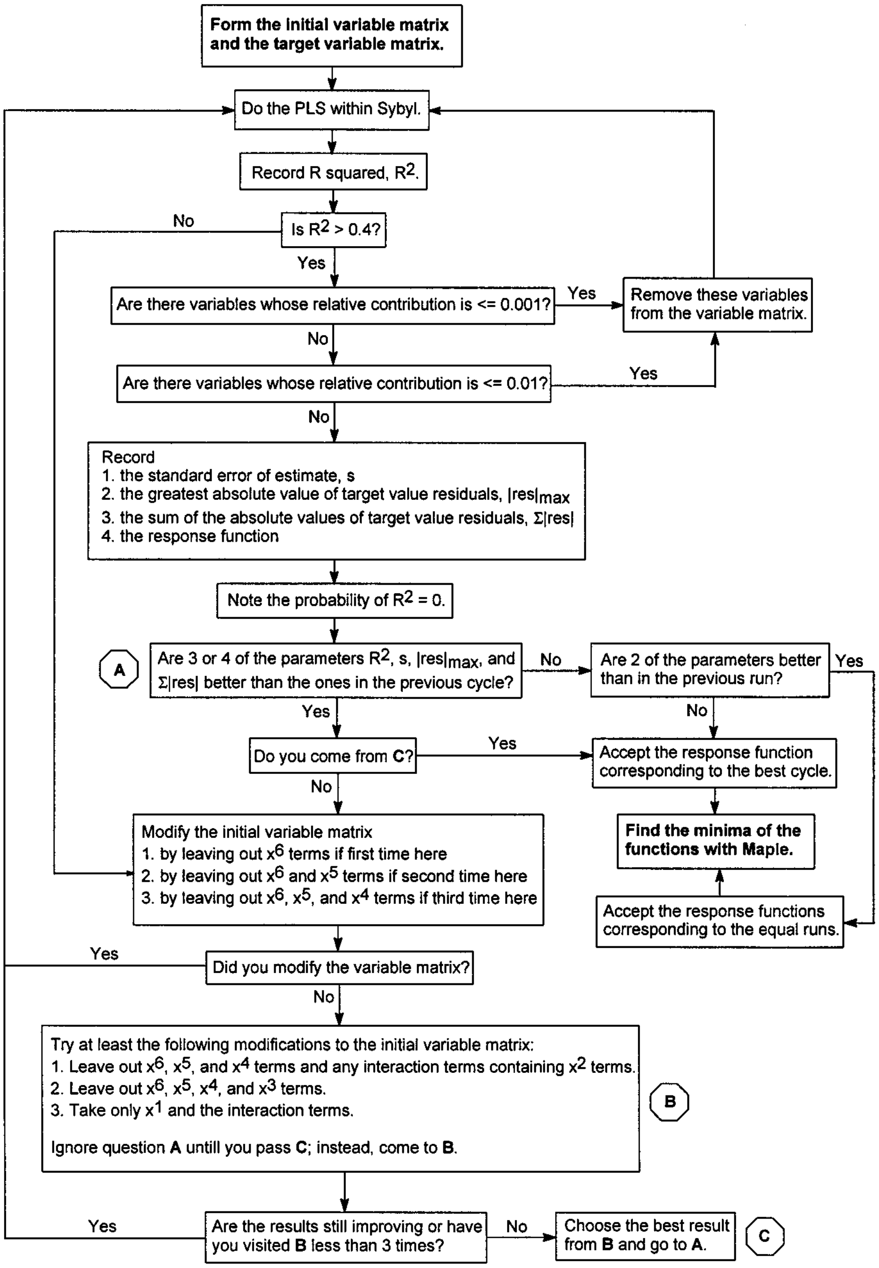

Form the initial variable matrix and the target variable matrix.
Do the PLS within Sybyl.
Record R squared, R2.
Is R2 > 0.4?
No
Yes
Are there variables whose relative contribution is <= 0.001?
Yes
Remove these variables from the variable matrix.
No
Are there variables whose relative contribution is <= 0.01?
Yes
No
Record
1. the standard error of estimate, s
2. the greatest absolute value of target value residuals, |res|max
3. the sum of the absolute values of target value residuals, Σ|res|
4. the response function
Note the probability of R2 = 0.
A
Are 3 or 4 of the parameters R2, s, |res|max, and Σ|res| better than the ones in the previous cycle?
No
Are 2 of the parameters better than in the previous run?
Yes
No
Yes
Do you come from C?
Yes
Accept the response function corresponding to the best cycle.
No
Modify the initial variable matrix
1. by leaving out x6 terms if first time here
2. by leaving out x6 and x5 terms if second time here
3. by leaving out x6, x5, and x4 terms if third time here
Find the minima of the functions with Maple.
Accept the response functions corresponding to the equal runs.
Yes
Did you modify the variable matrix?
No
Try at least the following modifications to the initial variable matrix:
1. Leave out x6, x5, and x4 terms and any interaction terms containing x2 terms.
2. Leave out x6, x5, x4, and x3 terms.
3. Take only x1 and the interaction terms.
Ignore question A untill you pass C; instead, come to B.
B
Yes
Are the results still improving or have you visited B less than 3 times?
No
Choose the best result from B and go to A.
C

Figure 4. The flow chart for the PLS analysis.

Resources. The experiments were run randomly on three SGI workstations; time comparisons were made on a 250 MHz R4400 Indigo2 Extreme.

RESULTS AND DISCUSSION

Parameter Refinement. Thirteen variables, as shown in Table 1, and two dummies were used in the first PBD cycle, the most important main effects can be seen in Table 2. It is noteworthy that both evaluation criteria, *i.e.*, the best force field score and the *rmsd* value associated with it, identified the same two parameters as the most favorable; the third ranking dummy variable receives a much lower main effect but is still indicative of some hidden interactions. The very drawback of the method is that the calculated main effects of the variables are confounded by and indistinguishable from two-factor and higher order interactions.[21] PBD has, however, been shown[22] to be superior as compared to a fractional factorial[37] and a random balance experiment.[38] There was also a dummy variable with considerable significance among the least favorable variables. The evaluation criteria scattered significance differently among these parameters; however, they agreed on the two most unfavorable parameters. Thus, four variables, *bump_maximum* (the maximum number of close contacts between the ligand and the protein to be allowed for an orientation to be optimized), *simplex_con-*

Table 2. The Results of the first PBD. m(ff) and m(*rmsd*) Are the Main Effects with the Best Force Field Score and the *rmsd* of the Best Scoring Orientation, Respectively, as Response

parameter	m(ff)	parameter	m(*rmsd*)
bump_maximum	−10.40	*bump_maximum*	−2.10
simplex_convergence	−6.13	*simplex_convergence*	−1.76
dummy	−3.75	dummy	−0.74
simplex_initial_rotation	3.35	dummy	0.85
distance_tolerance	3.47	*ligand_binsize*	1.60
simplex_restart	5.74	*interpolate*	1.70
dummy	6.16	*simplex_restart*	1.80
nodes_maximum/minimum	10.43	*nodes_maximum/minimum*	2.19

Table 3. Parameter Space and the Results of the first CCD, Test Points Were Scattered Evenly over the Range of the Variable

			best by		
parameter	min	max	PLS	matrix algebra	systematic searching
bump_maximum	0	4	4	1	4
simplex_convergence	0.2	0.6	0.2	0.2	0.2
simplex_restart	0.6	1.4	0.6	1.4	1.4
nodes_maximum/minimum	4	8	4	4	6
force field score[a]			−25.20	−25.19	−25.27
rmsd			0.48	0.46	0.48
run time (s)			2046	730	8336
predicted force field score			−28.67	−26.65	
greatest absolute residual			0.257	0.900	
sum of absolute residuals			2.285	5.051	

[a] Not comparable with Table 4.

vergence (convergence criterion within a simplex), *simplex_restart* (restart criterion for successive simplicies), and *nodes_maximum/minimum*, on which both criteria could agree, were chosen for the CCD; the other variables were set at their default values as shown in Table 1.

Since the main effects extracted from the PBD give a hint on the direction and magnitude of the change in response as one moves from (−) to (+), it is logical to stretch the ranges of the variables in the CCD beyond the ones used in the PBD experiments. For *bump_maximum*, the direction chosen was the only one available; for *nodes_maximum/minimum* on the other hand, the allowable range of 4−8 conveniently forms a five-point parameter space. It should, however, be mentioned that PBD correctly identified advantageous—although not necessarily the best—directions for these variables despite the limitations in choosing their ranges. The range for *simplex_convergence* was very conservatively set at 0.2−0.6, but one could easily argue that a range of, *e.g.*, 0.1−0.5, as suggested by the PBD, would have been more sensible. This view is corroborated by the results of the CCD: a minimum was found at one of the corners of the test space, as shown in Table 3, with a value of 0.2, *i.e.*, the lower boundary of the range, for *simplex_convergence*; this value was further lowered during the final refinement as can be seen from Table 1. The range of *simplex_restart* was expanded in the direction suggested by the PBD. Table 3 shows that this parameter receives the value of either boundary depending on the method used to analyze the results of the CCD. Therefore one cannot unambiguously state whether the PBD correctly recognized the favorable direction since either way will produce good values. The minima found by both the PLS and the full second-order polynomial model produced a slightly worse force field score than the best one seen in the CCD (−25.23). The run time of the best analysis was, however, over 7500 s, which was regarded as too high a price for such a small improvement in response.

It was observed during the second cycle that poor results were achieved if the dielectric function of the AMBER[39] force field was set distance independent. Therefore, a distance dependent function was always used. There was no need to calculate the main effects in the PBD since one of the test runs provided an *rmsd* value clearly superior to the others (0.38 Å against 0.45 Å or worse); however, the analysis time was still over 13 min (780 s). The parameter values of this run (*radmax* was increased from 4.0 to 5.0 to allow larger spheres in the shallow binding pocket), as shown in Table 1, were accepted as a part of the refined parameter set.

The second CCD was performed to explore the effect of five variables identified as important in the manual of DOCK3.5.[1] The *ligand_binsize* (the width in Å of the "bins" into which the ligand atom−atom distances are divided) and *ligand_overlap* (the overlap in Å between successive bins) were set equal to *receptor_binsize* (bin size for the spheres) and *receptor_overlap*, respectively,[5] which then reduces the number of optimizable variables down to three; the parameter space is shown in Table 4. Setting the overlaps dependent on the bin sizes clearly violates the principle of independent variables. By definition, however, these parameters cannot be regarded as independent even if not linked in this way. Two response functions were obtained for this CCD by the PLS method; one, however, had a much more beneficial minimum, which also turned out to be the global minimum, as judged by the corresponding dockings: an *rmsd* of 0.42 Å and a force field score of −26.15 against 0.54 Å and −25.63, respectively. The major problem with the obtained parameters was the CPU time of analysis, which was over 10 min for the 4000 orientations considered. Therefore, some additional testing with the bin size and overlap values was done; *simplex_convergence* and *simplex_restart* were adjusted as well to achieve very stringent criteria for the simplex convergence. The final parameter values are shown in Table 1, and they give an *rmsd* of 0.44 Å (score −26.14) in 7 min. It is obviously still too much for screening large databases. However, taking into account the stringency of simplex convergence, these parameters would be quite appropriate for a small set of ligands; 100 CPU s per ligand have been given as a reasonable example in the literature.[5]

Finally, it should be borne in mind that X-ray structures are approximations to a certain extent and that DOCK3.5[1] uses the rigid-body approach. The experimental reference used for comparison is a molecular model fitted onto an electron density map; an uncertainty of 0.15−0.2 Å for the position of a heavy atom can be typically expected for an X-ray structure with a resolution of 2 Å. Thus, looking for the optimal *rmsd* value may actually be meaningless. Indeed, in this study the orientations possessing the lowest force field scorings did not have the smallest *rmsd* deviations that were seen.

PLS Performance. The PLS performance can be judged by the results in Tables 3 and 4. It is seen from Table 3 that neither the PLS method nor the full second-order polynomial model can reproduce the best force field score achieved by systematic searching, but the difference in scores is very small. In the end, *bump_maximum* received a value of 5, and therefore the PLS method can be considered to have produced a more realistic value for this variable. Also, both *simplex_convergence* and *simplex_restart* ultimately received smaller values than the ones in Table 3. It is noteworthy that both methods used to analyze the results of the CCD found a minimum at one of the edges of the parameter space, which are—including the vertices—largely unexplored areas in a central composite design. Therefore, one should proceed with caution in such a case. On another occasion, the PLS method finds the global minimum recognized by systematic searching as can be seen from Table 4. In both test series, the force field scoring, which

Table 4. Parameter Space and the Results of the Second CCD, Test Points Were Scattered Evenly over the Range of the Variable

			best by		
parameter	min	max	PLS	matrix algebra	systematic searching
binsize	0.5	0.9	0.9	0.9	0.9
overlap	0.0**binsize*	0.4**binsize*	0.0**binsize*	0.0**binsize*	0.0**binsize*
bump_maximum	3	7	5	3	5
force field score[a]			−26.15	−25.69	−26.15
rmsd			0.42	0.39	0.42
time (s)			752	231	752
predicted force field score			−26.84	−27.28	
greatest absolute residual			0.436	0.222	
sum of absolute residuals			1.964	1.253	

[a] Not comparable with Table 3.

is the criterion used to judge the ligands in a database search, is the same or better with the PLS method than with the full second-order polynomial model. On both occasions the *rmsd* values are practically identical. The price to be paid for the possible enhanced performance is the run time, which is approximately three times higher for the PLS method.

It should be noted that one must be very careful with introducing third- and higher-order terms into the response function as they might easily help in fitting the polynom to the measured data without having any actual meaning. In our tests, however, the x^5 and x^6 terms always dropped out since they clearly worsened the results. Tables 3 and 4 show that mixed results were achieved as the predictive power of the response functions was compared. Based on this limited data set, it seems that the force field score is predicted with less accuracy if the residuals are small, *i.e.*, the model is better fitted to the data points. This observation must not be taken as a general rule; it probably applies only to a complex response surface which could be anticipated, *e.g.*, in the case of molecular interaction energies. Therefore, we can only recommend the use of higher-order terms if the resulting response function is *not* used as such for predicting responses by inter- or extrapolation but rather for finding the minima of the response surface within the parameter space. Also, a word of caution is appropriate: the full second-order polynomial model may fail as easily in a complex environment.

Fast Parameters. The best values for bin sizes, overlaps, and the matching distance tolerance were found to be 0.4, 0.3, and 0.7 (= 0.4 + 0.3), respectively. These parameters yielded an orientation with an *rmsd* of 0.43 Å and a score of −25.86 in just 16 s. Figure 5 shows that this orientation is slightly rotated when compared to the one with our refined parameters; the visual ranking of the two bottom orientations in Figure 5 is, however, very difficult, but they are clearly distinguishable from the orientation with the default parameters. The reduction in analysis time, albeit at the cost of a slightly worse scoring, is effected by two factors. First, the number of orientations minimized was only 1/19 of those with our refined parameters. On the other hand, with our refined parameters it is obvious that throughout the test series a considerable number of the orientations use *simplex_iterations* (maximum number of iterations of simplex minimization to be performed per orientation) minimization cycles (results not shown). Using 10 CPU s per ligand as a guideline for large databases,[5] this parametrization step could in comparable cases be realized in one working day with reasonable CPU time access.

Docking Experiment. Our docking experiment shows, as seen in Figure 6, that considerable rearrangement takes place when the ligands docked with the fast parameters (set A) are redocked with our refined parameters (set B). Among the first 50 ligands in set B, there are 11 ligands that ranked worse than 50 in set A; corresponding values for larger subsets are 17 out of 100, 16 out of 150, and 14 out of 200. Although the test set of 500 molecules is small and the top ranking molecules are not necessarily good ligands for the PKC subunit, we anticipate a similar phenomenon to be present if the results of an extensive database search were redocked.[12] Therefore, we suggest that large databases (in the order of 100 000 molecules) be docked with a parameter set allowing a fast screen, and 2000−5000 top scoring molecules then be redocked with a refined parameter set with more stringent conversion criteria to rearrange the ligands in this subset. Then possibly only 100−200 top scoring ligands[2] would require visual inspection to group these ligands into subsets of structurally similar molecules, whose representatives would be tested *in vitro*. (It is important to realize that a database like the ACD-3D[36] contains a plethora of very similar structures, which inevitably receive similar scorings when docked.) Smaller databases would naturally be directly docked with a refined parameter set.

Parameter Generality. The refined parameter set was tested with three other protein−ligand complexes. The known ligand orientations in 5ldh,[27] 3dfr,[28] and 1vid[29] were reproduced with varying accuracy. The *rmsd* value for the docked NAD-lactate in 5ldh was 13.5 Å, which is indicative of a totally different orientation. The crystal orientations of methotrexate in 3dfr (0.47 Å) and 3,5-dinitrocatechol in 1vid (0.33 Å) were reproduced correctly in 196 and 1260 s, respectively. The latter run time is obviously too long to be of practical value. It must be emphasized that the refined parameter set presented here is not suitable for general use since inaccurate orientations or considerable oversampling are possible. Instead, DOCK should be reparametrized for each individual docking problem.

CONCLUSIONS

Using DOCK as an example, we have shown that the Plackett−Burman and central composite designs followed by a PLS analysis are well suited for optimizing the parameters of computer programs in molecular modeling. The partial least squares method was found to be very useful in resolving the form of the response functions after the CCD. Conventionally, one first decides on the form of the response

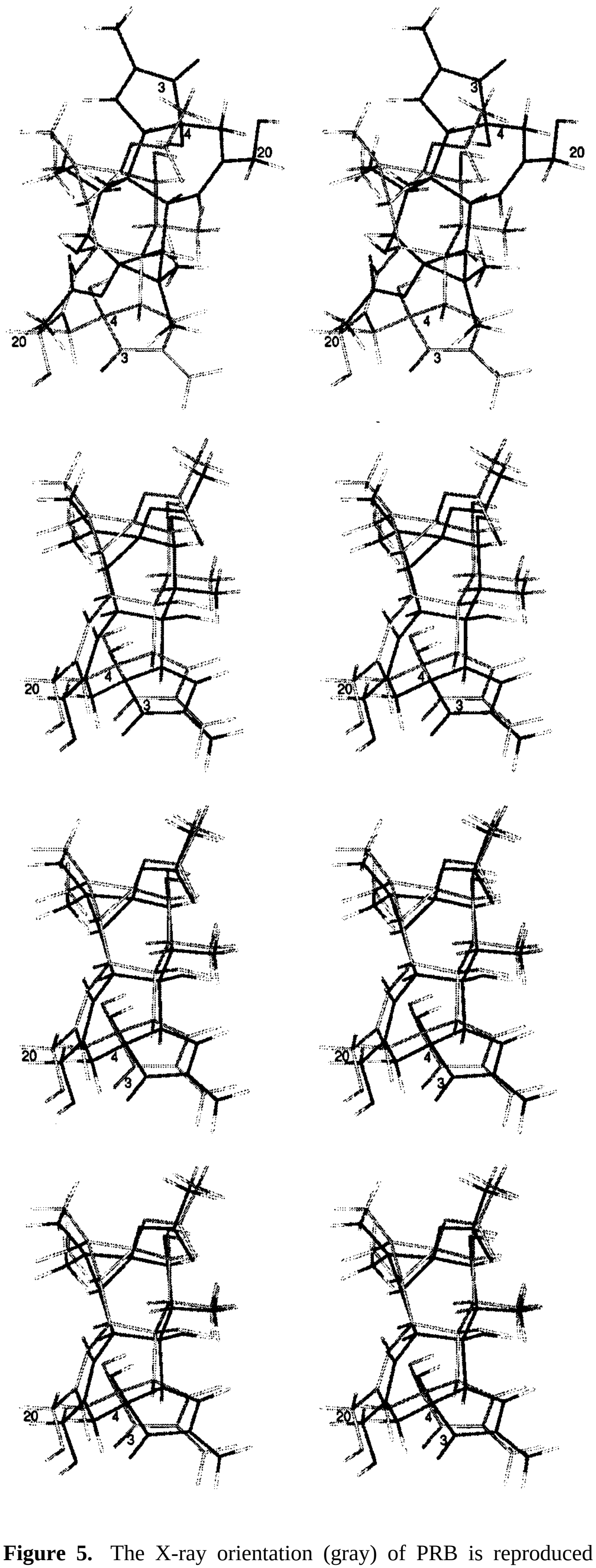


Figure 5. The X-ray orientation (gray) of PRB is reproduced (black) with varying accuracy depending on the parameter set used. The default parameters produce a best scoring orientation with an *rmsd* of 6.77 Å without minimization (top) and 0.61 Å with minimization (middle top). This value is 0.44 and 0.43 Å for our refined parameters (middle bottom) and the fast parameters (bottom), respectively. The carbonyl group on C-3 and the hydroxyl groups on C-4 and C-20 form hydrogen bonds with the protein.[26] The images were produced with InsightII.[40]

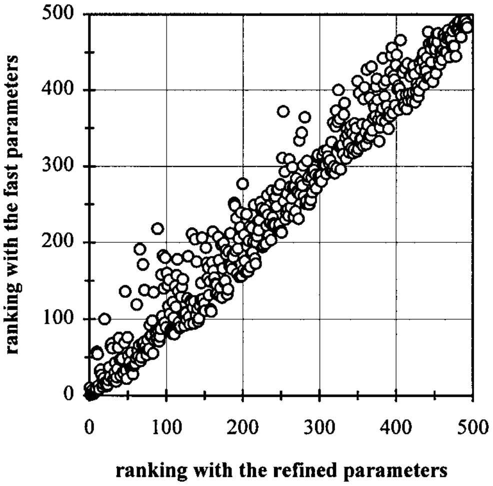


Figure 6. Choosing one or the other parameter set effects rearrangement of the ligand rankings.

function and then solves the values of its constants with the aid of matrix algebra. In our method, the correct form, *i.e.*, the number and the nature of its terms, and the values of the constants are resolved simultaneously in a stepwise manner. This approach could be useful for finding the minima/maxima in other similar optimizations with a large number of variables and a potentially complex response surface.

A set of parameter values was found that produces an orientation for the ligand with an interaction "energy" at the global minimum. The docking time, however, was prohibitively long for screening very large ligand databases. Therefore, a stepwise procedure is suggested consisting of (1) the docking of the database with a faster, but slightly less accurate set of parameters, (2) the redocking of a few thousand best scoring ligands from step 1 with a refined set of parameters, (3) the clustering of a few hundred best scoring ligands of step 2 into groups of structurally similar compounds, and (4) the testing of the representatives of the clusters in step 3. The refined parameter set presented here for 1ptr is, however, not suitable for general use, and we urge other workers to optimize the parameters for their particular problem and ligand.

ACKNOWLEDGMENT

This work was partly made possible by the financial support of the Technology Development Centre Finland (TEKES). The authors would also like to thank M. Sc. Janne Ravantti for his invaluable assistance in the realm of computers.

REFERENCES AND NOTES

(1) DOCK 3.5; UCSF: San Francisco, CA, 1995.

(2) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078−1082.

(3) Li, R.; Chen, X.; Gong, B.; Selzer, P. M.; Li, Z.; Davidson, E.; Kurzban, G.; Miller, R. E.; Nuzum, E. O.; McKerrow, J. H.; Fletterick, R. J.; Gillmor, S. A.; Craik, C. S.; Kuntz, I. D.; Cohen, F. E.; Kenyon, G. L. Structure-Based Design of Parasitic Protease Inhibitors. *Bioorg Med. Chem.* **1996**, *4*, 1421−1427.

(4) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(5) Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: On-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 123−132.

(6) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397.

(7) Meng, E. C.; Shoichet, B. K., Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.

(8) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.

(9) Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. Orientational Sampling and Rigid-Body Minimization in Molecular Docking. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 266−278.

(10) DesJarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; de Montellano, P. O.; Furth, P. S.; Alvarez, J. C.; DeCamp, D. L.; Babé, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 6644−6648.

(11) Shoichet, B. K.; Stroud, R. M.; Santi, D. V.; Kuntz, I. D.; Perry, K. M. Structure-Based Discovery of Inhibitors of Thymidylate Synthase. *Science* **1993**, *259*, 1445−1448.

(12) Bodian, D. L.; Yamasaki, R. B.; Buswell, R. L.; Stearns, J. F.; White, J. M.; Kuntz, I. D. Inhibition of the Fusion-Inducing Conformational Change of Influenza Hemagglutinin by Benzoquinones and Hydroquinones. *Biochemistry* **1993**, *32*, 2967−2978.

(13) Ring, C. S.; Sun, E.; McKerrow, J. H.; Lee, G. K.; Rosenthal, P. J.; Kuntz, I. D.; Cohen, F. E. Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3583−3587.

(14) Rutenber, E.; Fauman, E. B.; Keenan, R. J.; Fong, S.; Furth, P. S.; Ortiz de Montellano, P. R.; Meng, E.; Kuntz, I. D.; DeCamp, D. L.; Salto, R.; Rosé, J. R.; Craik, C. S.; Stroud, R. M. Structure of a Nonpeptide Inhibitor Complexed with HIV-1 Protease. *J. Biol. Chem.* **1993**, *268*, 15343−15346.

(15) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.

(16) Box, G. E. P.; Wilson, K. B. On the Experimental Attainment of Optimum Conditions. *J. R. Statist. Soc.* **1951**, *13*, 1−45.

(17) Plackett, R. L.; Burman, J. P. The design of optimum multifactorial experiments. *Biometrika* **1946**, *33*, 305−325.

(18) Read, D. R. The design of chemical experiments. *Biometrics* **1954**, *10*, 1−15.

(19) Palasota, J. A.; Deming, S. N. Central Composite Experimental Designs. Applied to Chemical Systems. *J. Chem. Educ.* **1992**, *69*, 560−563.

(20) Svoboda, V. Search for optimal eluent composition for isocratic liquid column chromatography. *J. Chromatogr.* **1980**, *201*, 241−252.

(21) Stowe, R. A.; Mayer, R. P. Efficient screening of process variables. *Ind. Eng. Chem.* **1966**, *58*, 36−40.

(22) Williams, K. R. Comparing screening designs. *Ind. Eng. Chem.* **1963**, *55*, 29−32.

(23) Routh, M. W.; Swartz, P. A.; Denton, M. B. Performance of the Super Modified Simplex. *Anal. Chem.* **1977**, *49*, 1422−1428.

(24) Betteridge, D.; Wade, A. P.; Howard, A. G. Reflections on the modified simplex − II. *Talanta* **1985**, *32*, 723−734.

(25) Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *J. Chemom.* **1994**, *8*, 111−125.

(26) Zhang, G.; Kazanietz, M. G.; Blumberg, P. M.; Hurley, J. H. Crystal Structure of the Cys2 Activator-Binding Domain of Protein Kinase Cδ in Complex with Phorbol Ester. *Cell* **1995**, *81*, 917−924.

(27) Grau, U. M.; Trommer, W. E.; Rossmann, M. G. Structure of the active ternary complex of pig heart lactate dehydrogenase with S-lac-NAD at 2.7 Å resolution. *J. Mol. Biol.* **1981**, *151*, 289−307.

(28) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal structures of Escherichia coli and Lactobacillus casei dihydrofolate reductase refined at 1.7 Å resolution. I. General features and binding. *J. Biol. Chem.* **1982**, *257*, 13650−13662.

(29) Vidgren, J.; Svensson, L. A.; Liljas, A. Crystal structure of catechol O-methyltransferase. *Nature* **1994**, *368*, 354−358.

(30) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(31) Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic Acids. *Science* **1983**, *221*, 709−713.

(32) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape. *J. Med. Chem.* **1986**, *29*, 2149−2153.

(33) Sybyl 6.3; Tripos, Inc.: St. Louis, MO, 1996.

(34) Shoichet, B. K.; Kuntz, I. D. Protein Docking and Complementarity. *J. Mol. Biol.* **1991**, *221*, 327−346.

(35) Maple V Release 4; Waterloo Maple Inc.: Waterloo, ON, 1996.

(36) Available Chemicals Directory 94.1; MDL Information Systems, Inc.: San Leandro, CA, 1994.

(37) Davies, O. L. In *The Design and Analysis of Industrial Experiments*; Oliver and Boyd: London, 1954.

(38) Satterthwaite, F. E. Random Balance Experimentation. *Technometrics* **1959**, *1*, 111−137.

(39) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr.; Weiner, P. A. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

(40) InsightII 95.0; Biosym/MSI: San Diego, CA, 1995.

CI9801825