SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization

Shay Bar-Haim,* Ayelet Aharon, Tal Ben-Moshe, Yael Marantz, and Hanoch Senderowitz

Epix Pharmaceuticals Ltd., 3 Hayetzira Street, Ramat Gan 52521, Israel

Received September 16, 2008

Identifying active compounds (hits) that bind to biological targets of pharmaceutical relevance is the cornerstone of drug design efforts. Structure based virtual screening, namely, the *in silico* evaluation of binding energies and geometries between a protein and its putative ligands, has emerged over the past few years as a promising approach in this field. The success of the method relies on the availability of reliable 3-dimensional (3D) structures of the target protein and its candidate ligands (the screening library), a reliable docking method that can fit the different ligands into the protein's binding site, and an accurate scoring function that can rank the resulting binding modes in accord with their binding affinities. This last requirement is arguably the most difficult to meet due to the complexity of the binding process. A potential solution to this so-called scoring problem is the usage of multiple scoring functions in an approach known as consensus scoring. Several consensus scoring methods were suggested in the literature and have generally demonstrated an improved ranking of screening libraries relative to individual scoring functions. Nevertheless, current consensus scoring strategies suffer from several shortcomings, in particular, strong dependence on the initial parameters and an incomplete treatment of inactive compounds. In this work we present a new consensus scoring algorithm (SeleX-Consensus Scoring abbreviated to SeleX-CS) specifically designed to address these limitations: (i) A subset of the initial set of the scoring functions is allowed to form the consensus score, and this subset is optimized via a Monte Carlo/Simulated Annealing procedure. (ii) Rank redundancy between the members of the screening library is removed. (iii) The method explicitly considers the presence of inactive compounds. The new algorithm was applied to the ranking of screening libraries targeting two G-protein coupled receptors (GPCR). Excellent enrichment factors were obtained in both cases: For the cannabinoid receptor 1 (CB1), SeleX-CS outperformed the best single score and afforded an enrichment factor of 41 at 1% of the screening library compared with the best single score value of 15 (GOLD_Fitness). For the chemokine receptor type 2 (CCR2) SeleX-CS afforded an enrichment factor of 72 (again at 1% of the screening library) once more outperforming any single score (enrichment factor of 20 by G_SCORE). Moreover, SeleX-CS demonstrated success rates of 67% (CCR2) and 73% (CB1) when applied to ranking an external test set. In both cases, the new algorithm also afforded good derichment of inactive compounds (i.e., the ability to push inactive compounds to the bottom of the ranked library). The method was then extended to rank a lead optimization series targeting the Kv4.3 potassium ion channel, resulting in a Spearman's correlation coefficient, $\rho = 0.63$ ($n = 40$), between the SeleX-CS-based rank and the actual pKi values. These results suggest that SeleX-CS is a powerful method for ranking screening libraries in the lead discovery phase and also merits consideration as a lead optimization tool.

INTRODUCTION

In silico or virtual screening has become a common practice in current computer aided screening efforts.[1] The method is used to reduce both the time and the cost involved in identifying hit compounds which bind to biological targets of pharmaceutical importance. One of the approaches frequently used in this field is structure based screening. The success of structure based *in silico* screening heavily relies on the availability of reliable 3D structures of both the target protein and the compounds comprising the screening library (the set of compounds to be screened against the protein's binding site). While the generation of the 3D conformations of druglike compounds from their 2D representations can be readily accomplished through the usage of commercially available software products such as Corina[2] or Concord,[3] obtaining reliable 3D structures of the target proteins may pose a challenge. In favorable cases, the crystal structure of the target protein might be available through the PDB.[4] In other cases, a model of the target protein has to be generated using a variety of computational techniques utilizing either an available structure of a homologue protein (*e.g.,* homology modeling)[5−7] or the protein's sequence (e.g., *de novo* methods).[8−11]

An important class of proteins of high pharmaceutical relevance are G protein coupled receptors (GPCRs). Members of this family are involved in many signaling pathways and have been targeted for treating multiple pathologies including neurological disorders, cardiovascular diseases, metabolic diseases, inflammation, and pain. Indeed in a recent review, it was reported that over 25% of FDA-approved drugs target GPCRs.[12]

* Corresponding author phone: +972-3-612-8590; fax: +972-3-612-8528; e-mail: sbar-haim@epixpharma.com.

10.1021/ci800335j CCC: $40.75
Published on Web 02/20/2009

Unfortunately, the crystal structures of most GPCRs (except that of bovine rhodopsin,[13] β_2 adrenergic receptor,[14] turkey β_1 adrenergic receptor,[15] and opsin[16,17]) are not available due to technical challenges thereby hampering structure-based drug discovery efforts of GPCRs-related therapeutics. One approach which has been successful in overcoming this difficulty is the PREDICT program,[8-11] which is a nonhomology based method for modeling the 3D structure of the trans-membrane domain of GPCRs from their amino acid sequences. This approach has been successfully applied to *in silico* screening of GPCR targets.[9]

Once the 3D structures of both the target protein and the screening compounds are available, virtual screening proceeds by docking the screening library into the protein's binding site using one of the available docking algorithms and by the concomitant evaluation of the interaction of the binding modes (poses) using a scoring function. DOCK,[18] GOLD,[19] LigandFit,[20] FlexX,[21] and Glide[22] are among the most widely used protein−ligand docking algorithms, but many more docking tools have been described in the literature.[23,24] Once docking is completed, the resulting poses can be further evaluated by additional scoring functions. Ideally, the scoring function should quantify the binding free energy of the ligand to the protein. An accurate calculation of binding free energies requires the evaluation of many complex energy terms, including solvation energy, electrostatic energy, VdW interactions, hydrophobic interactions, and various entropy terms. These calculations are time-consuming and can be applied only to small sets of ligands.[25] However for screening large databases of virtual compounds it is essential to evaluate the ligand-protein binding free energy at a very rapid pace. These conflicting requirements force most scoring functions to use simplified energy expressions instead of rigorous calculations of binding free energies, with different scoring functions highlighting different properties (see method section for some examples). This (over)simplification inevitably leads to significant inaccuracies, and, indeed, the scoring problem is largely considered to be as of yet unresolved.

The performance of a given scoring function (and in fact the entire screening process) is evaluated by its ability to rank-order the compounds in the screening library so that active compounds will rise to the top of the list. If this could be achieved, then biological testing of only the best scoring compounds in the library should be sufficient for identifying active hits. In order to rank the relative performance of the scoring functions (or of that of consensus scoring, see below) a quantitative measure is required. One such measure is the enrichment factor, which is calculated as the percent of active compounds found at the top X% of the library divided by the percent of the screening library that is being analyzed (X%). For example if 50% of the active compounds were found among the top 2% ranking compounds of the screening library, the enrichment factor would be 25-fold (50%/2%=25). Random selection is equivalent to an enrichment factor of 1.

A good scoring function not only should enrich the population of active compounds within the top portion of the library but also should reduce the population of inactive compounds at this same region. We term this feature "derichment". Derichment is calculated in a similar way to enrichment with the goal to have derichment factor <1.

A potential solution to the scoring problem is the usage of multiple scoring functions in an approach known as consensus scoring. Consensus scoring consists of scoring the binding modes resulting from the docking simulations with multiple scoring functions and then identifying those ligands that score well in some relevant subset of these functions. The scientific rationale behind this approach is that if indeed different scoring functions highlight different aspects of ligand-protein interactions (albeit approximately), a proper combination of these functions will result in a better description of the binding process.

Several consensus scoring strategies have been described in the literature[1,26-35] and mainly differ in the way the individual scores are combined into a consensus. The "standard" rank-by-vote[35] (also known as vote-by-percent[33]) method uses all scoring functions to rank all members of the screening library, and a vote is cast by each scoring function in favor of a particular compound if its score falls above a predefined, constant threshold. The consensus score of a compound is simply the sum of its votes, and the different compounds are ranked according to their consensus scores. If the individual scoring functions are well correlated with binding affinities, active compounds are expected to surface to the top of the list. Thus, results obtained with the rank-by-vote paradigm depend on the initial selection of scoring functions and threshold values. Moreover this method can lead to multiple compounds having identical ranks, if applied to a screening library that was scored with few scoring functions (For N scoring functions the consensus score is an integer between 0 and N.). This situation makes the selection of candidates for biological assaying difficult. Furthermore, since most scoring functions cannot distinguish between active and inactive compounds when part of a "tight" SAR series, active compounds surfacing to the top of the list often "drag" with them inactive compounds as well. The rank-by-vote method is implemented for example in Cerius^2's Consensus Scoring and in Sybyl's Cscore modules. Other consensus scoring methods are 'rank-by-number' which ranks each compound according to its averaged normalized score values[33] and 'rank-by-rank' which sorts the screening library based on the average rank of each compound as determined by the individual scoring functions.

Several works discussing consensus scoring have been published over the past few years covering both the theoretical aspects[35] and the applications[1,26,30-35] of this strategy. Charifson et al.[30] have used a simplified version of the rank-by-vote strategy to score a set of ligands against several protein targets of pharmaceutical relevance and demonstrated a consistent improvement in the false positive rate with only a minor reduction in the rate of true positives. Clark et al.[31] have demonstrated that the usage of several consensus scoring strategies can reliably pick the most active compounds from within a set of ligands of diverse activity docked into a variety of target proteins. Finally, Krovat and Langer[32] have tested the consensus scoring module within Cerius[2,36] and found that the usage of a quadruple consensus scoring of the LigScore2, PLP1, PLP2, and JAIN scoring functions resulted in a hit rate of 90% in the top 1.4% of a randomly generated screening library that was characterized by a MW between 450 and 600 and docked into the binding site of the aspartic protease rennin. This hit rate was higher than those obtained with the individual scoring functions.

Why does consensus scoring work? Wang and Wang[35] have suggested a theoretical explanation based on statistical arguments. According to this explanation, the success of consensus scoring results from a mere cancelation of random errors upon summation of multiple scores. The main assumption underlying this argument is that errors embedded in "good" scoring functions are random. While this assumption is certainly true in the ideal case, contemporary scoring functions are far from being perfect. Indeed, the neglect of certain energy terms as well as the approximate treatment of others may not necessarily result in random errors. Moreover, if the statistical argument is extended, then a near-perfect ranking of a screening library could be obtained simply by adding up a very large number of scoring functions. Nevertheless Yang et al.[29] have shown that although the accuracy of consensus scoring can be improved by increasing the number of scoring methods (2- and 3-combination in their example), combining all scoring methods (i.e., 5-combination) did not lead to the best performance.

An alternative explanation emphasizes the complementarity of the different scoring functions in describing the energy terms responsible for the ligand-protein free energy of binding. Consider for example a hypothetical binding site that is characterized by two types of interactions, H-bonding and electrostatic. Consider also two real scoring function, GOLD[19] and LigScore,[37] that have scored a set of potential ligands docked into the hypothetical site. Since GOLD lacks an electrostatic term and LigScore lacks an explicit H-bonding term, taking one of these scoring functions alone will probably result in erroneous ranking. On the other hand, combining these function through the mechanism of consensus scoring will most likely result in improved ranking (limited of course by the accuracy of the calculated energy terms).

Despite the successes of consensus scoring strategies, several important questions have remained unanswered. In particular, how to select the best combination of individual scores, how to filter out scoring functions that do not adequately describe the binding process for the target at hand, and how to incorporate the information about inactive compounds into the consensus scoring process.

In this work we address these questions by presenting a novel consensus scoring algorithm which combines a comprehensive search of the consensus score space with a unique ranking strategy. The new algorithm was successfully applied to the virtual screening of two screening libraries targeting two different proteins. It was then utilized in a lead optimization program of a third target, to rank a Structure Activity Relationship (SAR) series of compounds with encouraging results.

METHODS

Preparation of Protein Models. Models for the transmembrane domain of the CCR2 Chemokine and the CB1 Cannabinoid receptor, both belonging to the G-Protein Coupled Receptor (GPCR) family, were generated using the PREDICT algorithm as detailed elsewhere.[8−11] The final models were refined in the presence of suitable reference ligands in their binding sites through molecular dynamics (MD) simulations using CHARMm,[38] in accord with the PREDICT refinement protocol.[8−11] The model for the Kv4.3 potassium ion channel was generated by homology modeling using the Methanobacterium thermoautotrophicum potassium channel (Mthk) as a template (PDB code 1LNQ) with the MODELER module of Accelrys's Discovery Studio.[39] The model was refined in the presence of a suitable reference compound in its binding site through MD simulations using CHARMm.

Preparation of Screening Libraries. Epix's screening library consists of over 4,500,000 unique compounds stored in their 1D forms in an Oracle database. This library was obtained from over 30 catalogues of vendors and is continuously updated. The screening library was processed prior to the docking simulation as detailed earlier.[9] Briefly, structures were converted into their 3D form using CONCORD v6.8[3] and were assigned SYBYL atom types[40] and Gasteiger charges.[41] Multiple conformations were generated using Epix's HYPERION software in combination with CONFORT v6.8.[42] The full screening library was filtered based on the characteristics of the binding site (charged, polar, hydrophobic residues) and on the properties of known ligands (molecular weight, druglike properties) affording a focused screening library.

Docking. The focused screening libraries (a different one for each target protein) were docked into the proteins' binding sites using the genetic algorithm-based docking program GOLD v2.2[19] for CB1 and CCR2 and DOCK v4.0[18] for Kv4.3.

Conformation Selection. The resulting binding modes (poses) were filtered by means of Epix's Binding Mode Analysis algorithm in order to remove erroneous poses and to select the pose which best fits the binding site. The BMA/Sitewise algorithm ranks the different binding modes according to the proximity of functional groups within the ligand, to their appropriate counterparts within the binding site which are considered to be important for binding. In doing so, BMA/Sitewise applies a "chemist's insight" to the analysis of the docking results.

Scoring Functions. A brief description of the scoring functions used in this work is provided in the Supporting Information. More details are available in the original publications.[18−20,25,37,40,43−49] The selected poses (one for each ligand) were rescored using a per-case subset of scoring function from the list (see Tables S1, S3, and S5 in the Supporting Information for more details).

Consensus Scoring: SeleX-CS. SeleX-CS is an extension of the rank-by-vote consensus scoring approach, featuring some important additions:

a. Initial parameters: The main new features in SeleX-CS are as follows: (i) Only a subset of the scoring functions is allowed to cast votes in favor of the members of the screening library. This subset is optimized through the Monte Carlo/Simulated Annealing (MC/SA) procedure (see below). (ii) Thresholds are allowed to vary on a per-scoring function basis, and their values are also optimized through MC/SA yielding individually optimized thresholds for each scoring function.

b. Actual ranking: In contrast to "standard" rank-by-vote, each compound is ranked according to two values, a primary, rank-by-vote value and a secondary, rank-by-number value, thereby greatly extending the range of different scores and consequently reducing ranking redundancy.

c. Inactive compounds: SeleX-CS explicitly considers the presence of inactive compounds (if information of such compounds is available) and tries to "push" them toward the bottom of the ranked list.

SeleX-CS accepts as input a study table containing information on the compounds comprising the focused screening library (actives, inactives, and unknowns). This information consists of an activity tag ("1" for actives, "−1" for inactives, and "0" for unknowns) and score values obtained by the different scoring functions. These scores are normalized to the [0−1] range, and the resulting nominal values are divided into 128 bins and are arranged so that the highest value always corresponds to the best scoring compound. Based on a randomly selected subset of the scoring functions, each compound is assigned a $CS_{compound}$ value which is subsequently used to rank the entire study table. A target function, *f*, is then applied to the ranked library to calculate a SeleX-CS value ($CS_{library}$). $CS_{library}$ is calculated from the number of active compounds at the top of the ranked list and from the number of inactive compounds at its bottom. This value is optimized through an MC/SA procedure by considering different subsets of scoring functions and threshold values. See the "target function" section for a detailed example. Once optimized, the best $CS_{library}$ value defines that combination of scoring functions and threshold values which should be used for the selection of active compounds. There is no theoretical limit on the number of scoring functions that can be used by SeleX-CS, nor is there a limit on the granularity of the threshold values.

Monte Carlo/Simulated Annealing. Monte Carlo/Simulated Annealing (MC/SA)[50] combines two widely used approaches, SA and MC, into a general procedure for the global optimization (minimization or maximization) of any target function $f(\vec{x})$ whose value (in our case $CS_{library}$) depends on several parameters x_i. In the current implementation, *f* is the target function whose value is maximized, and its parameters (x_i) are both the scores and their threshold values. This method follows a Simulated Annealing predefined cooling schedule within a temperature range and runs an MC simulation at each temperature.

Starting from a high temperature (T_{max}) the algorithm runs a certain number of MC steps, N_{MC} (either predefined or until convergence has been reached). Each step consists of applying a number of perturbations (N_p) to the current subset of scores and threshold values (termed configuration). A perturbation may add, remove, or replace a score or modify the value of a threshold. The new configuration is evaluated relative to its predecessor according to the Metropolis acceptance criteria (eq 1)

$$p = \min[1, e^{-\frac{\Delta f}{RT}}] \qquad (1)$$

where *R* is a constant, *T* is the current temperature, and Δf is the difference between two subsequent evaluations of *f*. In the current implementation of the metropolis criteria we use the term ($\Delta f/RT$) rather than $-(\Delta f/RT)$ since we attempt to maximize the value of the target function.

The temperature is then reduced according to the SA cooling schedule, and the MC simulation is repeated. The procedure terminates once the temperature is reduced to below the lower temperature limit (T_{min}). In order to avoid trapping at local minima, the cooling cycle is repeated multiple times (see Figure 1).

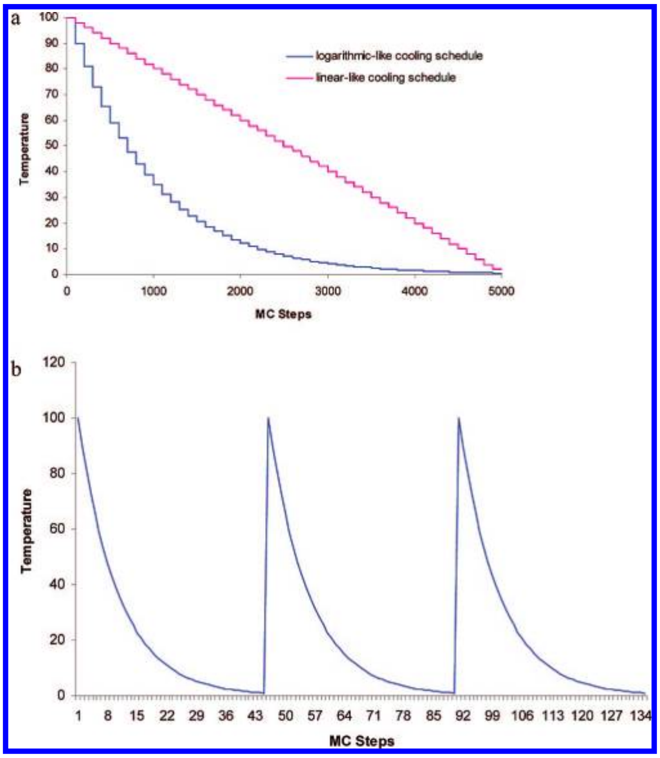

Figure 1. Different cooling schedules available in SeleX-CS. (a) Depiction of a linear versus a logarithmic-like cooling schedule, with a constant number of steps at each temperature (default setup). (b) A combined logarithmic-like/saw-tooth cooling schedule implemented so as to avoid trapping in local minima.

Calculation of $CS_{compound}$. The calculation of $CS_{compound}$ proceeds as follows:

1. Select, at random, a subset of scores ({s}) and their corresponding thresholds ({*c*}).

2. Assign to each compound a primary ($Score_{primary}$) and a secondary ($Score_{secondary}$) score. The primary score is calculated according to the rank-by-vote method whereby, if the value of the scoring function for a compound (s_i) exceeds the threshold (c_i), it is awarded one vote. This procedure is repeated for all of the scoring functions (i) included in the subset. Thus for a selected subset scores of size N, the primary value ranges between 0 and N.

$$Score_{primary} = \sum_{i=1}^{scoring\_functions\_subset} S_i \qquad S_i = \begin{cases} 1 & s_i > c_i \\ 0 & s_i < c_i \end{cases} \qquad (2)$$

3. $Score_{secondary}$ for a compound consists of a summation of the normalized scores (n_i) which were allowed to cast a vote in favor of this compound. As noted above each score is normalized to receive a value between 0 and 127. Thus for a selected scores subset of size N, this secondary value ranges between 0 and Nx127.

$$Score_{secondary} = \sum_{i=1}^{scoring\_functions\_subset} n_i \qquad n_i \in \{0...127\} \qquad (3)$$

Consider, for example, Table 1 which lists threshold values (c_i) selected for a subset of 4 hypothetical scores (1, 3, 4, and 8). The table further lists the rank, according to the 4 scores (n_i), of a hypothetical compound (compound A) which is part of a 1000 compound screening library. Since the ranks of compound A exceed the threshold selected for scores 4 and 8 only, the primary consensus score ($Score_{primary}$) for this compound is 2. Further, the hypothetical nominal values calculated for compound A by the fourth and eighth scoring functions (score4 and score8, respectively), when considered relative to the 1000 compounds comprising the screening library (n_i), correspond to bins 119 (score4) and 98 (score8). The $Score_{secondary}$ value of compound A is therefore 119 + 98 = 217.

4. Sort the compounds comprising the screening library according to $Score_{primary}$ and then $Score_{secondary}$, thereby generating a ranked list.

Calculation of CS$_{library}$. $CS_{library}$ is calculated by the target function *f* from the ranked screening library where ranking is performed according to the individual $CS_{compound}$ values.

As noted above, the purpose of the *in silico* screening process is to select a small subset of compounds from within a screening library to be subjected to biological testing. If the screening library could be ordered so that known active compounds are located near its top (active region) whereas inactive compounds, near its bottom (inactive region), then unknown compounds residing in the active region may indeed be active. Thus, the target function in the present implementation should strive to maximize the presence of known active compounds within the top portion of the ranked library and the presence of inactive compounds (if present) within its bottom portion. The exact definition of the top and bottom portions ($N_{top\_prec'}$, N_{bot_prec}) is user-defined and is expressed as percentage of the size of the screening library.

The value of the target function, $CS_{library}$, of the ranked list is calculated as the weighted sum of 3 terms:

a. The number of known active compounds that appear at the top portion of the library ($count_{active}$; see eq 4).

b. The spread of the known active compounds which appears beyond the top portion of the library ($spread_{active}$)

$$spread_{active} = \sum_{j=1}^{active\_compounds} \frac{1}{(Rank_j - Rank_t)^n} \qquad (4)$$

where $Rank_j$ is the position of active compound *j* in the ranked list defined in section 3 above, and $Rank_t$ is the position in the ranked list defining the lower boundary of the library's top portion. In our example, for a library of 1000 compounds and a user-defined top portion of 10 (1%), $Rank_t = 10$. *n* defines the steepness of the spread function. As *n* decreases, the value of $spread_{active}$ increases and with it the tendency of known active compounds to concentrate just beyond the library's upper portion.

c. The spread of the known inactive compounds calculated from the bottom of the library ($spread_{inactive}$)

$$spread_{inactive} = \sum_{j=1}^{inactive\_compounds} \frac{1}{(Rank_j - Rank_t)^n} \qquad (5)$$

where $Rank_j$ is the position of inactive compound *j* in the ranked list defined in section 3 above, and $Rank_t$ is the position equivalent to the bottom of the library, in our example 1000. *n* has the same meaning as in eq 4.

Thus, the value of the spread functions increases as known active compounds are closer to the user defined top portion of the library and/or when known inactive compounds are closer to the bottom of the library. User defined weights are allocated to these 3 terms.

The final value of the target function, $CS_{library}$, is calculated as the weighted sum of these 3 terms:

$$CS_{library} = w1 * (count_{active} + spread_{active}) + w2 * spread_{inactive} \qquad (6)$$

The Final Model. The application of SeleX-CS to any specific data set results in an optimally selected set of scoring functions and their individual corresponding threshold values which maximizes $CS_{library}$. We term this combination a "model". This model uniquely defines the optimal ranked library from which enrichment factors are calculated (as defined above). A similar calculation could in principle be performed to quantify the derichment of inactive compounds. However, in this work, we chose to treat the inactive derichment in a quantitative manner. Thus, for each of the 3 examples considered in this work (see the Results section) we provide both an enrichment graph and a derichment graph from which the performance of the SeleX-CS can be readily evaluated.

SeleX-CS was coded and implemented in Object Oriented C++ and compiled with Microsoft Visual C++ 6.0. A typical SeleX-CS run takes approximately 6 h on a PentiumIII 2.8 GHz processor.

Discovery Studio - Consensus Scoring. Each data set was also tested in the Consensus Score protocol implemented in Discovery Studio 2.1[39] under the Receptor−Ligand Interactions category. The parameters used were the default ones provided by the vendor.

RESULTS

In this work we present results of three different SeleX-CS runs performed on different protein targets and at different stages of the drug discovery process. Two of the proteins are family A GPCRs (Cannabinoid receptor 1 (CB1) and Chemokine receptor type 2 (CCR2)) and were considered at the hit discovery phase while the third is an ion channel

Table 1. Calculation of $Score_{primary}$ and $Score_{secondary}$ for a Hypothetical Compound According to the SeleX-CS Algorithm[a]

	Scores subset				
	Score1	Score3	Score4	Score8	value
selected threshold (expressed as %) of screening library (c_i)	10%	50%	1%	7%	
selected threshold (expressed as number of compounds given a score of 1)	100	500	10	70	
rank of compound A, out of 1000 compounds (n_i)	150	600	6	50	
contribution to Primary value ($Score_{primary}$)	no	no	yes	yes	2
contribution to Secondary value ($Score_{secondary}$) (assumed values in the [0−127] range)	0	0	119	98	217

[a] See text for more details.

(the Kv4.3 potassium channel) considered at the lead optimization phase.

The selection of the individual scoring functions per target as input for the SeleX-CS runs was done by the program's scientists based on their understanding of the relevant receptor−ligand interactions and the performance of the individual scores.

Cannabinoid Receptor 1 (CB1). The data set for this receptor consisted of 214 known active compounds and 15 known inactive compounds together with a focused screening library of 16,765 focused compounds.[51] These were docked into the protein's PREDICT-generated model using GOLD, analyzed with BMA/Sitewise and the selected binding modes scored using 33 different scoring functions (see Table S1 in the Supporting Information). The raw screening data are provided in the Supporting Information (CB1_Data.txt).

The SeleX-CS run was configured to select score subsets $\{s\} = 3$ to 10 and to allow a maximum number of perturbations, $N_p = 3$. The top and bottom portions of the library were defined as $N_{top\_prec} = N_{bot\_prec} = 1\%$. This definition approximately reflects the number of compounds to be sent to biological assaying. The cooling schedule for the MC/SA consisted of 300 temperature cycles between $T_{max} = 1000$ K and $T_{min} = 1$ K, each with $N_{MC} = 1000$. The temperature was reduced in a logarithmic fashion leading to faster cooling at high temperatures and to slower cooling at low temperatures (see Figure 1). Equal weights (w) were assigned to the different terms of the target function (eq 6), and the power (n) used for the spread function was set to 1 (eqs 4 and 5). The best SeleX-CS model is presented in Table S2 of the Supporting Information.

The average enrichment factor obtained with the individual scoring functions at 1% of the library was 1 (i.e., no better than random), with GOLD_fitness providing the best enrichment factor of 15. Combining some of these scores through the SeleX-CS mechanism led to a marked improvement as evident by the enrichment factor of 41. In contrast, the consensus scoring mechanism as implemented in Discovery Studio was unable to enrich the top 1% of the library beyond a random selection (i.e., an enrichment factor of 1). Figure 2 compares the enrichment and derichment graphs obtained with SeleX-CS, Gold_fitness, and with Discovery Studio. Since we chose to optimize for $N_{top\_prec} = 1\%$, it is not surprising that the best performance and the most pronounced improvement in comparison with Gold score and Discovery Studio are observed at low %library values. Figure 2 also demonstrates that the derichment obtained with SeleX-CS is far superior to that obtained with Gold_score or Discovery Studio.

In order to further validate the performance of the method we divided the data into a training set and a test set. For this purpose, 25 known active compounds and 5 known inactive compounds were removed from the data set (by classifying them as unknowns), and their ranks were predicted by the algorithm together with the other members of the screening library. Thus these compounds constitute a valid independent test set. Eighteen of the 25 active compounds were ranked toward the top of the list and 4 out of the 5 inactive compounds, toward its bottom for a total success rate of 73%. Since all test compounds were taken from SAR series and were therefore structurally similar to the active compounds used to generate the model, these results reinforce our previous suggestion that scoring functions as well as consensus scoring strategies are pressed hard when facing the challenge of distinguishing between structurally similar active and inactive compounds. Still we were happy to see that 72% of the active compounds were properly ranked, suggesting that using SeleX-CS for compound selection will result in actual hits.

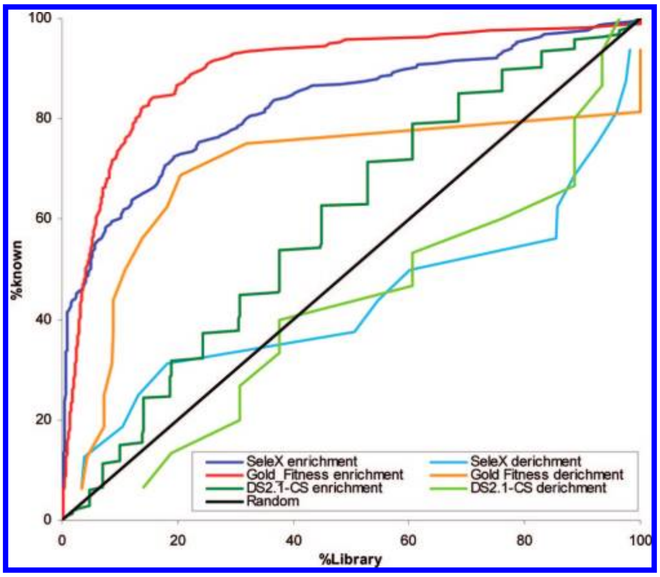

Figure 2. Enrichment/derichment graph for the best SeleX-CS model, the Discovery Studio consensus model, and the best individual score (GOLD_fitness) for the CB1 receptor. Enrichment factors, at 1% of the library, of 41, 15, and 1 were obtained for SeleX-CS, GOLD_fitness, and Discovery Studio, respectively. The line at 45° corresponds to the probability of identifying active or inactive compounds at random (i.e., no enrichment or derichment). Thus, a curve above this line corresponds to enrichment and a curve below the line, to derichment. Moreover, the steeper the curve near its origin (X=0), the better is the enrichment. In a similar manner, the steeper the curve near the X=100 point, the better is the derichment. The improved inactive derichment obtained with the consensus score model in comparison with GOLD is clearly evident.

Chemokine Receptor Type 2 (CCR2). The data set for this receptor consisted of 36 known active compounds and 34 known inactive compounds together with 31,764 focused compounds. These were docked into the protein's binding site using GOLD, analyzed with BMA/Sitewise and the selected conformations scored with a set of 12 scoring functions (see Table S3 in the Supporting Information). The raw data are provided in the Supporting Information (CCR2_Data.txt).

The SeleX-CS run was configured to select score subsets $\{s\} = 2$ to 8 and to allow a maximum number of perturbations $N_P = 3$. The top and bottom portions of the library were once more defined as $N_{top\_prec} = N_{bot\_prec} = 1\%$. The cooling schedule for the MC/SA consisted of 70 temperature cycles between $T_{max} = 1000$ K and $T_{min} = 1$ K, each with $N_{MC} = 1000$. The temperature was reduced in a logarithmic fashion, similar to the CB1 case. The weights in eq 6, $w1$ and $w2$, were assigned values of 0.7 and 0.3, respectively, and the power (n) used for the spread function was set to 3. The best SeleX-CS model is presented in Table S4 of the Supporting Information.

Once more, the results obtained with our method (an enrichment factor of 71 at 1% library) outperformed those obtained with either the individual scoring functions (average

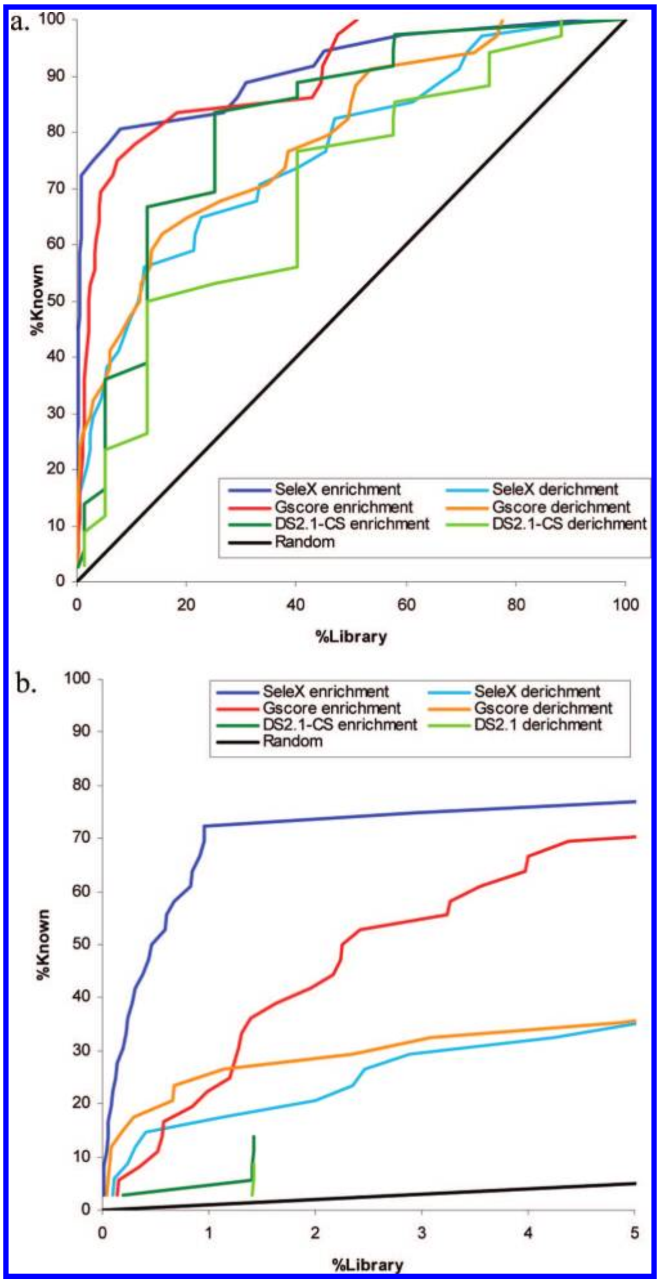

Figure 3. Enrichment/derichment graph for the best SeleX-CS model, the Discovery Studio model, and the best individual score (G_Score), for the CCR2 receptor. Enrichment factors, at 1% of the library, of 72, 20, and 5 were obtained for SeleX-CS, G_Score, and Discovery Studio, respectively. Once more, the improved inactive derichment obtained with the consensus score model in comparison to GOLD is clearly evident. Part b focuses on the region defined by %lib = [0%−5%].

enrichment factor of 10; best enrichment factor of 20 obtained with G_Score) or with Discovery Studio (enrichment factor of 5), and the same is true for the derichment (see Figure 3a,b).

As before, we divided the data into a training set and a test set. For this purpose, 9 known active compounds and 9 known inactive compounds were removed from the data set, and their ranks were predicted by the algorithm. Five of the 9 active compounds were ranked toward the top of the list and 7 out of the 9 inactive compounds, toward its bottom for a total success rate of 67%, similar to the results obtained for CB1.

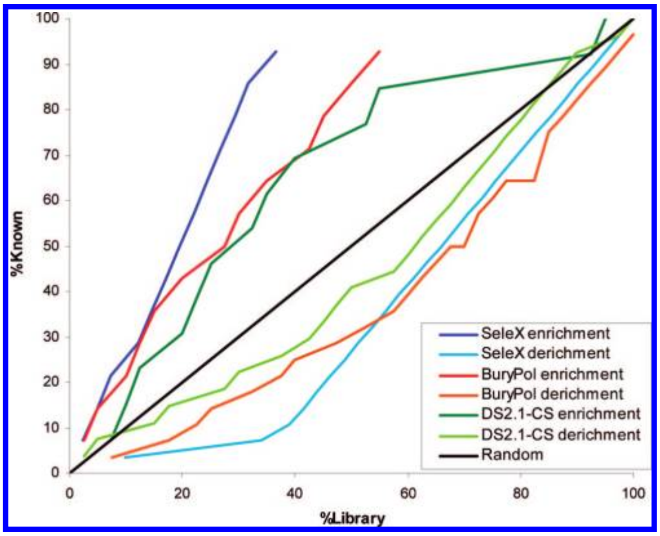

Figure 4. Enrichment/derichment graph for the best SeleX-CS model, the Discovery Studio model, and the best individual score (BuryPol2_CFF_LigScore2) for the Kv4.3 ion channel. Enrichment factors, at 32% of the library, of 2, 1.6, and 1.5 were obtained for SeleX-CS, BuryPol2_CFF_LigScore2, and Discovery Studio, respectively.

Kv4.3 Potassium Channel. In this project we attempted to expand the boundaries of SeleX-CS beyond its original purpose (*i.e., in silico* screening) and into the lead optimization phase in order to see whether the method can reliably predict the rank ordering (activity wise) of the next cohort of compounds to be synthesized. SeleX-CS was run as before, but the data set in this case consisted of known active and inactive compounds only (13 and 27, respectively) with no focused screening library. The set of 40 compounds was docked into the protein's binding site using DOCK and analyzed with BMA/Sitewise, and the selected binding modes were scored with a set of 19 scoring functions (see Table S5 in the Supporting Information). Raw data are provided in the Supporting Information (KV43_Data.txt).

The SeleX-CS run was configured to select score subsets $\{s\} = 2$ to 4 (to guard against over fitting) and to allow a maximum number of perturbations $N_P = 2$. The top and bottom portions of the library were defined as $N_{top\_prec} = N_{bot\_prec} = 32.5\%$. The logarithmic-like cooling schedule for the MC/SA consisted of 70 temperature cycles between $T_{max} = 1000$ K and $T_{min} = 1$ K, each of $N_{MC} = 1000$. Equal weights (w) were assigned to the different terms of the target function, and the power (n) used for the spread functions was set to 1. The best SeleX-CS model is presented in Table S6 of the Supporting Information.

The enrichment obtained with this model, at 50% of the library, was 2, while the average enrichment of the individual scores was approximately 0.8 with the best enrichment of 1.6 obtained with BuryPol^2_CFF_LigScore2. Discovery Studio showed an enrichment of 1.5. The derichment obtained with SeleX-CS was better than that obtained with BuryPol2_CFF_LigScore^2 as can be seen in Figure 4.

The rank ordering for the compounds, as obtained from the SeleX-CS run, was compared to that obtained from the pKi values using Spearman's correlation[52] leading to a ρ value of 0.63. For comparison, the Spearman correlation between the BuryPol^2_CFF_LigScore2-derived rank and the pKi-derived rank was slightly lower at 0.57.

DISCUSSION

Scoring a data set of diverse compounds docked into a protein's binding site in order to discriminate between active and inactive ones is probably the most difficult problem facing current virtual screening efforts. This so-called scoring problem is manifested in the myriad of scoring functions, which have been developed over the past few years. These functions differ from one another both in the method by which they were derived (*e.g.,* force field based, empirical and knowledge based) and in the way they treat the energy terms, which describe the binding process. Since each scoring function, by virtue of its parameters, defines its unique applicability domain, it is highly unlikely that a single function will be able to correctly describe all types of ligand−protein interactions or be applicable in all cases. Indeed, experience has shown that different scoring functions work well in different cases.

The consensus scoring paradigm is designed to increase our confidence in the scoring results. The main idea underlying all consensus scoring strategies is that the higher the number of scoring functions that rank a certain compound at a high place, the higher is the probability of that compound to be active. While a discussion still exists as to the preferred way of implementing this idea, most researchers accept its validity and agree that consensus scoring outperforms single scoring functions.

Our own implementation of the consensus scoring approach draws on previous ideas, in particular, the rank-by-vote and rank-by-number strategies but extends on them by incorporating a MC/SA mechanism for scores optimization, a nonredundant ranking strategy, and an implicit treatment of inactive compounds.

The method accepts as input a focused screening library, characterized by several scores. It then performs a MC/SA optimization in the score space in order to identify a unique combination of scores and their corresponding threshold values (a model) that maximize the number of active compounds at the top of the library and the number of inactive compounds at its bottom. This model provides as output a ranked list of the library members from which active enrichment and inactive derichment could be calculated in order to evaluate its performance. If deemed successful, the (unknown) compounds at the top of the ranked list are classified as active and sent to biological assaying. In the following we discuss the most important features of each step.

Which scoring functions should be presented to the algorithm? This problem is similar in nature to the descriptors selection problem in QSAR studies. Common practices in this field argue in favor of selecting a small subset of descriptors, which could be easily interpreted. We follow these recommendations by using a rather small set of scores as input to SeleX-CS. Furthermore, each set is tailored to the specific problem at hand and could be readily interpreted by looking at its components (i.e., constituting energy terms). In QSAR analysis, further filtration of descriptors is sometimes performed, for example, by selecting those descriptors, which on their own correlate best (R^2-wise) with the activity. This could be done for SeleX-CS as well, e.g., by considering only those scores, which provide the best enrichment on their own. While beneficial at times, this procedure, although feasible, is not always recommended. Two hypothetical scores, which by themselves lack the ability to differentiate between active and inactive compounds may well do so while working in tandem (see Figure 5 for a hypothetical example). Instead we favor an alternative approach in which we perform a MC/SA directed search in the combined scores/threshold space in search for the global maximum. This approach might be considered analogous to the stepwise regression or genetic algorithm (GA) approach often-employed in QSAR.

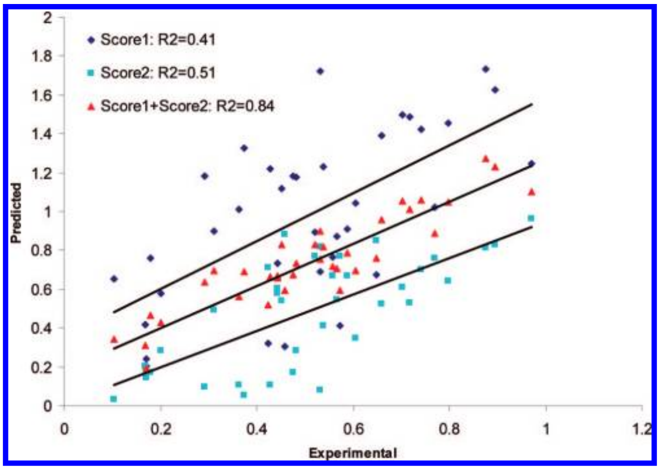

Figure 5. Two hypothetical scores, which by themselves provide only minimal correlation to the experimental values, are providing a much better correlation when working in tandem.

Limiting the number of scores presented to the algorithm may help preventing the chance correlation problem. In addition, care should be taken in order to avoid the problem of overfitting. To this end, we limit the number of scores, which are allowed to form the consensus at each MC step. The maximum number of scores can vary according to the size of the data set, with larger sets allowing for larger scores subsets, but in all cases should not be allowed to exceed a reasonable value.

Several parameters control the MC/SA simulation. The most important ones are the cooling schedule and the number of perturbations per MC step. Though we did not perform an exhaustive search of all possible combinations we did optimize these parameters in several test runs (data not shown).

Next we consider the nature of the target function. This was specifically designed to meet the ultimate goal of the screening process, namely, the discovery of at least one hit within that portion of the screening library, which will be sent for biological assaying. This portion is defined in our algorithm in terms of a percentile of the screening library and is termed N_{top_prec}. As outlined in the "Methods" section, the target function is comprised of 3 terms, describing, respectively, the number of active compounds within N_{top_prec}, the spread of active compounds just beyond N_{top_prec}, and the spread of inactive compounds toward the bottom of the library (N_{bot_prec}). The relative importance of these terms is governed by their weights (see eq 6) and by the value of n in eqs 4 and 5. Our default setup however ensures the dominancy of the first term. Consequently, a solution having one known active compound within N_{top_prec} and all other known actives spread far below it is preferred over a solution with all known active compounds ranked just below N_{top_prec}

but with no active compounds within it. Thus, it is appropriate to question the usage of the spread term (eqs 4 and 5). However, we reason that this term helps in distinguishing between solutions that are similar with respect to the presence of active compounds within N_{top_prec}. Clearly, it makes more sense to choose a model that on top of ranking known active compounds within N_{top_prec} has also resulted in a favorable distribution of active compounds just below it.

One of the major problems that traditionally haunts the hit discovery phase of drug development projects, whether performed through virtual or high throughput screening approaches, is the high rate of false positives. Consider for example a data set with 100,000 compounds, 10 of which are active. Even with a false negative rate as high as 50%, 5 active compounds could still be identified. On the other hand, a false positive rate as low as 1% will erroneously define 1000 inactive compounds as active. If financial resources are limited, for example, to screening only 300 compounds, there is a fair chance that none of the active compounds will be assayed at all.

This high rate of false positives at least partially results from the inability of current scoring functions to discriminate between structurally similar active and inactive compounds. Consequently, when active compounds are “surfacing” to the top of a ranked list, obtained either from a specific scoring function or by some consensus scoring mechanism, they almost inevitably “carry” with them inactive compounds as well.

In order to overcome this problem, our target function explicitly treats known inactive compounds (eq 5). This term, which is analogous to the spread term used for known active compounds, encourages inactive compounds to cluster toward the bottom of the ranked screening library.

When coming to evaluate the performance of SeleX-CS, it is therefore instructive to carefully analyze the enrichment/derichment graphs. The CB1 case is particularly encouraging. Figure 2 shows an excellent active enrichment within the top 1% (~170 compounds) of the library coupled with clear inactive derichment. Of the ~170 compounds included in this range, 41 are known actives and only 2 are known inactives. Figure 3a presents similar results for CCR2. At a first glance, these are less compelling than for CB1 as both active and inactive enrichments are observed. However, taking a closer look at the relevant portion of the screening library (Figure 3b) reveals that within the top 1% of the library the number of known active compounds (26) far exceeds that of inactive compounds (5). This trend continues within the top 5% of the library where the number of active and inactive compounds is 27 and 11, respectively.

From our own experience and from the experience of others[51] it is vitally important to evaluate a model’s performance by applying it to several cases that the model did not ‘see’ while being constructed. This set, termed external validation set or test set, can give more reliable information as to the future performance of the model than the training set (i.e., the set of compounds used for model generation).

To this end, we have evaluated both CB1 and CCR2 models by removing a portion of the known compounds (CB1: 25 active and 5 inactive compounds; CCR2: 9 active and 9 inactive compounds) in favor of test sets. These compounds were ranked, together with the rest of the screening library, by the respective models obtained with SeleX-CS. For the CB1 case, 18 of the 25 active compounds and 4 of the 5 inactive compounds were properly ranked for a total success rate of 73%. The corresponding numbers for CCR2 were 5 out of 9 for the active compounds and 7 out of 9 for the inactive compounds for a total success rate of 67%. These results, while not perfect, strongly suggest that using SeleX-CS for compound selection will result in actual hits. Due to the importance of external validation, it might be useful to incorporate the model’s performance on a test set as an additional term in the target function. This of course will require the generation of two test sets, one for model selection and the other for model validation. The usage of external validation as a criterion for model selection has been discussed by others.[53]

It is interesting to try and draw some inference from the threshold values obtained for the different scores in the final SeleX-CS model. Generally speaking, the threshold value for a given score is correlated with the discriminative power of this score. A high threshold value corresponds to a nondiscriminating score, whereas a low threshold value might lead to overfitting. In our studies we have let the cutoffs roam freely in the range of 0% to 100%, and indeed the resulting models presented threshold values covering this entire range. It is also interesting to note that in all cases, the scores that had the best individual performance showed up in the consensus score with a relatively low threshold value, indicating the importance of these scores.

The third example considered in this work, Kv4.3, is an attempt to utilize SeleX-CS as a tool for lead optimization. While lead optimization efforts would certainly benefit from a quantitative prediction of binding free energies, in many cases an accurate rank prediction (i.e., which is the more active compound) might suffice to move a program forward. As rank prediction is the natural working ground of SeleX-CS, we have calculated the Spearman’s correlation[52] between the results obtained from the best model generated by the algorithm and the actual pKi values. The result, $\rho = 0.63$, for a set of 40 compounds suggests a potential use of the method in lead optimization.

It is important to note that the development of any empirical computational tool for the early stages of lead optimization is hampered by the paucity of experimental data at this stage. Thus, it might be relevant to analyze the number of data points required to build a useful SeleX-CS model. Such an analysis is currently being considered in our group.

CONCLUSIONS

In this work we presented the development of a new consensus scoring algorithm and its application to the ranking of screening libraries targeting two GPCR proteins, CB1 and CCR2. The work was then extended to rank a smaller library of known compounds targeting a third protein, namely, the Kv4.3 potassium ion channel. In all cases, the consensus scoring mechanism provided good enrichment factors, much better than those obtained with the individual scoring functions or another consensus scoring software (i.e., Discovery Studio which was available to us). Moreover, the consensus scoring also afforded good derichment of inactive compounds.

SeleX-CS is based on known consensus scoring strategies, in particular the rank-by-vote and the rank-by-number.

However, it differs from them and from other common practices in this field in several important points: First, a subset of all scoring functions is considered, and this subset, together with its corresponding threshold values, is optimized via a Monte Carlo/Simulated Annealing procedure. Second, rank redundancy between the members of the screening library is removed by using a combination of a primary and a secondary score. Finally, the method explicitly considers the presence of inactive compounds.

SeleX-CS was developed independent of the nature of the docking algorithm and the scoring functions. Thus, we did not attempt to demonstrate that the docking and scoring methodologies employed throughout this work are better or worse than other alternatives. Rather we reason that if indeed the data presented to SeleX-CS contain a "real" biological signal, our method is more likely to reveal it than others. In this respect it is important to emphasize that the good results obtained with SeleX-CS do not diminish, in any way, the need for improved docking and scoring methodologies. Indeed the performance of our method is likely to improve upon such new developments.

SeleX-CS can be further developed in several ways. Perhaps the most obvious one is the replacement of the threshold values by a set of coefficients to be optimized through the MC/SA procedure. In this way it might be possible to replace the rank values calculated for the members of the screening library by actual binding constants. Work along these lines is currently being considered in our group.

ACKNOWLEDGMENT

The authors would like to thank Drs. Oren M. Becker, Boaz Inbal, Ori Kalid, Merav Fichman, Nili Schutz, Ms. Inbal Sela, and Mr. Erez Livneh for their contributions to the development and applications of the SeleX-CS algorithm.

Supporting Information Available: Final SeleX-CS models (scores and thresholds) for all examples considered in this work, together with the raw data used for model derivation. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44* (7), 1035–1042.

(2) *TSAR Corina*, Accelrys Inc.: San Diego, CA 92121, U.S.A., 1990.

(3) *CONCORD, 7.3*; Tripos Inc.: St. Louis, MO 63144, U.S.A., 2006.

(4) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(5) Fiser, A.; Do, R. K.; Sali, A. Modeling of loops in protein structures. *Protein Sci.* **2000**, *9* (9), 1753–1773.

(6) Marti-Renom, M. A.; Stuart, A.; Fiser, A.; Sãnchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291–325.

(7) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.

(8) Becker, O. M.; Dhanoa, D. S.; Marantz, Y.; Chen, D.; Shacham, S.; Cheruku, S.; Heifetz, A.; Mohanty, P.; Fichman, M.; Sharadendu, A.; Nudelman, R.; Kauffman, M.; Noiman, S. An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J. Med. Chem.* **2006**, *49* (11), 3116–3135.

(9) Becker, O. M.; Marantz, Y.; Shacham, S.; Inbal, B.; Heifetz, A.; Kalid, O.; Bar-Haim, S.; Warshaviak, D.; Fichman, M.; Noiman, S. G protein-coupled receptors: in silico drug discovery in 3D. *PNAS* **2004**, *101* (31), 51–86.

(10) Becker, O. M.; Shacham, S.; Marantz, Y.; Noiman, S. Modeling the 3D structure of GPCRs: advances and application to drug discovery. *Curr. Opin. Drug Discovery Dev.* **2003**, *6* (3), 353–361.

(11) Shacham, S.; Marantz, Y.; Bar-Haim, S.; Kalid, O.; Warshaviak, D.; Avisar, N.; Inbal, B.; Heifetz, A.; Fichman, M.; Topf, M.; Naor, Z.; Noiman, S.; Becker, O. M. PREDICT modeling and in-silico screening for G-protein coupled receptors. *Proteins* **2004**, *57* (1), 51–86.

(12) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5* (12), 993–996.

(13) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289* (5480), 733–734.

(14) Rasmussen, S. G. F.; Choi, H. J.; Rosenbaum, D. M.; Kobilka, T. S.; Thian, F. S.; Edwards, P. C.; Burghammer, M.; Ratnala, V. R. P.; Sanishvili, R.; Fischetti, R. F.; Schertler, G. F. X.; Weis, W. I.; Kobilka1, B. K. Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **2007**, *450*, 383–387.

(15) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486–491.

(16) Park, J. H.; Scheerer, P.; Hofmann, K. P.; Choe, H. W.; Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **2008**, *454*, 183–187.

(17) Scheerer, P.; Park, J. H.; Hildebrand, P. W.; Kim, Y. J.; Krauss, N.; Choe, H. W.; Hofmann, K. P.; Ernst, O. P. Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **2008**, *455*, 497–502.

(18) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15* (5), 411–428.

(19) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267* (3), 727–748.

(20) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21* (4), 289–307.

(21) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470–489.

(22) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.

(23) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **2006**, *65* (1), 15–26.

(24) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16* (3), 151–166.

(25) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate - DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

(26) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43* (25), 4759–4767.

(27) Teramoto, R.; Fukunishi, H. Consensus Scoring with Feature Selection for Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 288–295.

(28) Betzi, S.; Suhre, K.; Chetrit, B.; Guerlesquin, F.; Morelli, X. GFscore: A General Nonlinear Consensus Scoring Function for High-Throughput Docking. *J. Chem. Inf. Model.* **2006**, *46*, 1704–1712.

(29) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

(30) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42* (25), 5100–5109.

(31) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20* (4), 281–295.

(32) Krovat, E. M.; Langer, T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1123–1129.

(33) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J. Chem. Inf. Model.* **2006**, *46* (1), 380–391.

(34) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46* (12), 2287–2303.

(35) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1422–1426.

(36) *Cerius2, 4.11*; Accelrys Inc.: San Diego, CA 92121, U.S.A., 1990.

(37) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23* (5), 395–407.

(38) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.

(39) *Discovery Studio, 2.1*; Accelrys Inc.: San Diego, CA 92121, U.S.A., 2008.

(40) *SYBYL, 7.3*; Tripos Inc.: St. Louis, MO 63144, U.S.A., 2006.

(41) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. *Tetrahedron* **1980**, *36* (22), 3219–3228.

(42) *CONFORT, 7.2*; Tripos Inc.: St. Louis, MO 63144, U.S.A., 1998.

(43) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G., V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425–445.

(44) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2* (5), 317–324.

(45) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1998**, *17* (5), 490–519.

(46) Jain, A. N. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10* (5), 427–440.

(47) Mohamadi, F.; Richard, N. G. J.; Guida, W. C.; Liskamp, R. M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. MacroModel - an Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comput. Chem.* **1990**, *11*, 440.

(48) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791–804.

(49) Verkhivker, G. M.; Rejto, P. A.; Bouzida, D.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Gehlhaar, D. K.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Towards understanding the mechanisms of molecular recognition by computer simulations of ligand-protein interactions. *J. Mol. Recognit.* **1999**, *12* (6), 371–389.

(50) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(51) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793–806.

(52) Spearman, C. General Intelligence objectively determined and measured. *Am. J. Psych.* **1904**, *15*, 201–293.

(53) Golbraikh, A. S., M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2−4), 241–253.

CI800335J