A Simple Method to Improve the Odds in Finding 'Lead-Like' Compounds from Chemical Libraries

Kouhei Horio,[a] Hajime Muta,[a] Junichi Goto,[b] and Noriaki Hirayama*,[a]

[a] *Tokai University School of Medicine; Boseidai, Isehara, Kanagawa 259–1193, Japan: and* [b] *Computational Science Department, Science & Technology Systems Division, Ryoka Systems Inc.; 1–28–38 Shinkawa, Chuo-ku, Tokyo 104–0033, Japan.* Received November 20, 2006; accepted April 20, 2007; published online April 25, 2007

A simple method of virtual screening is proposed. This method uses only chemical characters calculated from two dimensional chemical structures. Local and global chemical characters are represented by molecular fingerprint and trait, respectively. The trait is a newly introduced concept in this paper and it is expressed by a set of two dimensional (2D) molecular descriptors. In this study, both the molecular fingerprint and the trait were used to represent drug-likeness of a group of molecules with a particular pharmacological activity. To learn about the molecular fingerprint and trait specific to a particular group of drugs, we used a database of drugs that are clinically used in Japan now. The molecular fingerprint and trait trained on these real drugs were used to predict drug-likeness of molecules in other chemical databases. In these chemical databases, an appreciable number of compounds that show the relevant pharmacological activity are contained. Some of these compounds are drugs clinically used abroad, but not in Japan. The prediction rate was judged by an enrichment factor. Despite the simplicity of the methodology, practical results were obtained. In the case of β-adrenergic blockers, the enrichment factor of 66 was attained and nearly 57% of active molecules in the chemical databases were successfully covered.

Key words virtual screening; drug-likeness; molecular descriptor; molecular fingerprint; trait

A large number of new technologies have been introduced into pharmaceutical research in recent years to provide information about potential drug targets and their disease association. In most drug discovery projects, however, we are still in the situation where one does not have any specific target in mind at the time when a set of molecules for screening is necessary. In this situation, we often use small number of active compounds as clues to discover better drug candidates. Since quite a large number of chemicals are commercially available, the discovery process usually starts from screening drug candidates in these chemical databases. In pharmaceutical companies, the proprietary compounds are also added to the target databases to be screened. A tremendous number of compounds is usually contained in such databases. Therefore, it is usually not realistic that even modern high-throughput screening technique can actually handle all of the compounds. If it is possible to reduce the number of compounds to be experimentally tested, it can make the entire process of searching cost effective and fast. Virtual screening can serve as a tool to suggest compounds to be tested experimentally. In this sense, virtual screening is a computational counterpart to high-throughput screening. As virtual screening is an entirely computer-based method, it can search a very large chemical space. Virtual screening is becoming an important screening technique because it is also very cost effective.

In this paper, a simple method of virtual screening is proposed. It is usually difficult to identify the specific conformation of a drug that is responsible to the biological activity. Accordingly, in many drug discovery projects, we have no alternate but to use only two dimensional chemical structures. Since chemical characters of drug molecules determine their drug-likeness, this method uses only chemical characters calculated from the two-dimensional chemical structures of the molecules. Local and global chemical characters are represented by molecular fingerprint and trait, respectively. The concept of molecular fingerprint is relatively well-established in the field of cheminformatics. As a molecular fingerprint is expressed by substructural information, the molecular fingerprint represents a local chemical character of a molecule. The trait, however, is newly introduced concept in this study and it can represent a global character of a set of molecules. A trait of a particular group of molecules is expressed by a combination of various two dimensional (2D) molecular descriptors. The 2D molecular descriptors only use the atoms and connection information of the molecule. The trait represents common global characters of these molecules. A drug-likeness profile of a set of drugs with a particular pharmacological activity can be expressed by a combination of the molecular fingerprint and the trait. To learn about drug-likeness of drugs with a specific pharmacological activity, we used a database of drugs that are clinically used in Japan (DCUJ) now. The number of independent drugs in DCUJ was 1170 when we undertook this study. Namely, DCUJ was used as a training dataset. The drug-likeness profile trained on these real drugs was employed to predict the specific drug-likeness of compounds in other chemical databases. In these chemical databases which consist of thousands of compounds, an appreciable number of compounds that show the relevant pharmacological activity are contained. Some of these compounds are drugs clinically used abroad, but not in Japan. The prediction rate was judged by an enrichment factor. Despite the simplicity of the methodology, reasonably practical results have been successfully obtained.

Experimental

Database About 1200 independent drug molecules are contained in DCAJ and these molecules were used to make training datasets. Drugs belonging to four categories *i.e.* anxiolytic agents, β-adrenergic blockers, ACE inhibitors and antibiotics were used to make the training dataset for each drug category. There are several chemical databases that contain many drugs and drug-like compounds. These databases can be used as test datasets for validation. Merck Index (Merck)[1] and Comprehensive Medicinal Chemistry (CMC)[2] contain many drug and drug-like compounds. The information of

* To whom correspondence should be addressed. e-mail: hirayama@is.icc.u-tokai.ac.jp

investigative new drugs is also useful to consider the drug-likeness of compounds. In this study a database of the investigative new drugs under clinical trials in Japan (IND)[3] is taken into account. We merged these three databases and removed the compounds contained in DCUJ and other redundant compounds to construct a test dataset consisting of 8277 unique compounds. The test dataset was used to evaluate the predictive ability of the method mentioned below.

Molecular Fingerprint Binary forms of chemical structures based on 2D chemical representation are called molecular fingerprints. Several methodologies exist for chemical binary representations. Molecular Design Limited (MDL) created a key based molecular fingerprint. This molecular fingerprint uses a pre-defined set of definitions and creates molecular fingerprints based on pattern matching of the structure to the defined "key" set.[4] This key based approach relies on the definitions to encapsulate the molecular descriptions. The keys were originally developed for the purpose of database substructure searching. Each "key" describes a small substructure consisting of about one to ten non-hydrogen atoms. These substructures can be used to characterize molecules. Molecular fingerprints also can be used to filter out compounds that do no meet certain criteria. In this study, the bit packed version of the 166 public MDL MACCS structural keys was used to represent molecular fingerprints.

Trait Molecular fingerprints only represent local molecular characters. Global characters of drugs are also highly important to determine drug-likeness of molecules. We have introduced a new concept of trait to express common global characters of a set of drugs that possess a similar pharmacological activity. To represent the trait of drugs of a particular pharmacological category, the distribution patterns of 2D descriptors were used. In this study, 124 descriptors as shown in Table 1 were used.

Most of the descriptors in Table 1 are described in ref. 5 and implemented in the software system MOE.[6] The asterisked descriptors are newly introduced for this study. Brief explanations for these descriptors are as follows:

(1) contents of atoms

a_wY: the weight percent of the atom type Y in a molecule. The *sp*, sp^2 and sp^3 hybridization states of atoms are expressed by sp, sp2 and sp3, respectively. Hali, Haro and Hact mean hydrogen atoms bonded to aliphatic carbon atoms, aromatic carbon atoms and other types of atoms, respectively.

(2) molecular dimension

'd' and 'r' are the largest and smallest values in the distance matrix, respectively. d/r is used to represent the relative shape of the molecule.

(3) bonds

b_xxxxR: number of bonds of bond type xxxx (including implicit hydrogens) divided by the number of all bonds (including implicit hydrogens). "ar" denotes aromatic bond.

b_count_mw: number of bonds (including implicit hydrogens) divided by molecular weight

b_heavy_mw: number of bonds between heavy atoms divided by molecular weight

(4) connectivity and shape index

chiN_mw: atomic connectivity index (order N) divided by molecular weight

chiNv_mw: atomic valence connectivity index (order N) divided by molecular weight.

chiN_Cmw: carbon connectivity index (order N) divided by molecular weight.

chiNv_Cmw: carbon valence connectivity index (order N) divided by molecular weight.

It is not easy to consider the correlations between the molecular descriptors and find the independent descriptors. For instance, molecular weight and volume are generally highly correlated, but it is obvious that there are not always correlated. Therefore in this study all molecular descriptors are treated independently. The ranges of the values of 124 molecular descriptors were calculated for a set of drugs of a particular category. Then the ranges that cover a certain percentage of the drugs were determined. A trait of a particular group of drugs is expressed by a set of these ranges of the 124 molecular descriptors.

Traits of several drug categories expressed by a small set of descriptors are shown in Table 2. In these examples, 14 typical molecular descriptors are used to represent global characters, *i.e.* traits. The range for each descriptor was determined in order to cover 80% of drugs with the same pharmacological activity incorporated in DCUJ. The range is characteristic of each drug category. The ranges of the descriptors observed for anxiolytic drugs are relatively narrow. For antibiotics, however, the ranges are appreciably wide. The ranges of the descriptors that cover 80% of all drugs ('drugs' in Table 2) in DCUJ are also noteworthy. They, all together, represent a drug-likeness profile of drugs in general. By use of 124 descriptors instead of 14, we can more appropriately define the trait of each drug category. The ranges of all 124 descriptors that cover 80% of all drugs in DCUJ are given Supplemental Table S1.

Table 1. 124 Descriptors Used to Express the Trait of Molecule

contents of atoms
a_wBr,* a_wCl,* a_wF,* a_wI,* a_wP,* a_wS,* a_wN,* a_wNsp,* a_wNsp2,* a_wNsp3,* a_wO,* a_wOsp2,* a_wOsp3,* a_wH,* a_wHali,* a_wHaro,* a_wHact,* a_wC,* a_wCsp,* a_wCsp2,* a_wCsp3*
grobal properties
weight, vdw_vol, density, d/r,* vdw_area, RingN (Oprea), SlogP, TPSA, SMR, apol, bpol, Fcharge
bonds
b_singleR,* b_doubleR,* b_tripleR,* b_arR,* b_count_mw,* b_heavy_mw*
pharmacophore features
a_acc, a_acid, a_base, a_don, a_hyd
information content
a_IC, a_ICM
bond rotation
b_1rotN, b_rigidN (Oprea), b_rotN, b_1rotR, b_rotR
connectivity and shape indices
chi0_mw,* chi1_mw,* chi0v_mw,* chi1v_mw,* chi0_C_mw,* chi1_C_mw,* chi0v_C_mw,* chi1v_C_mw,* KierA1, KierA2, KierA3, KierFlex, Zagreb
vertex
VAdjEq, VAdjMa.
adjacency and distance matrix descriptors
VDistEq, VdistMa, balabanJ
subdivided surface areas
SlogP_VSA0, SlogP_VSA1, SlogP_VSA2, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, SlogP_VSA6, SlogP_VSA7, SlogP_VSA8, SlogP_VSA9, SMR_VSA0, SMR_VSA1, SMR_VSA2, SMR_VSA3, SMR_VSA4, SMR_VSA5, SMR_VSA6, SMR_VSA7
surface areas
vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_hyd, vsa_other, vsa_pol
partial charge descriptors
PEOE_PC−, PEOE_PC+, PEOE_RPC−, PEOE_RPC+, PEOE_VSA_POS, PEOE_VSA_FPOS, PEOE_VSA_PPOS, PEOE_VSA_FPPOS, PEOE_VSA_NEG, PEOE_VSA_FNEG, PEOE_VSA_PNEG, PEOE_VSA_FPNEG, PEOE_VSA_POL, PEOE_VSA_FPOL, PEOE_VSA_HYD, PEOE_VSA_FHYD, PEOE_VSA+0, PEOE_VSA+1, PEOE_VSA+2, PEOE_VSA+3, PEOE_VSA+4, PEOE_VSA+5, PEOE_VSA+6, PEOE_VSA−0, PEOE_VSA−1, PEOE_VSA−2, PEOE_VSA−3, PEOE_VSA−4, PEOE_VSA−5, PEOE_VSA−6

Asterisked descriptors are made for the present study.

Drug-Likeness and Enrichment Enrichment procedure by use of molecular fingerprints was undertaken by the molecular fingerprint model function implemented in MOE. Molecular fingerprints calculated for drugs in the training dataset were used as the reference molecular fingerprints. The molecular fingerprints of drugs in the training dataset were compared to the molecular fingerprints of compounds in the test dataset. The similarity between all drugs in the training dataset and a compound in the test dataset was judged by the concordance between their molecular fingerprints. The compounds with molecular fingerprints that exceeded the specified similarity thresholds were regarded as compounds that should possess the relevant biological activity. The Tanimoto index[7] was used as a similarity metric and 80% was used as a specified similarity threshold in the present study. 80% was a good threshold that covers highly similar compounds.

Ranges of the values of 124 descriptors calculated for a group of compounds can characterize the common features of the group. Therefore the ranges that cover 80% of drugs with a particular pharmacological activity should well represent the trait of this family of drugs. The trait was used to judge a propensity of a compound for having the specific pharmacological activity. If a descriptor of a compound in a test dataset took a value within the range defined by the trait, a 1(hit) was assigned. If the value was out of the range, a 0 was assigned. The number of hits divided by 124 takes a continuous value from 0 to 1. The drug probability based on this value was used to find drug-like compounds. In this study the compounds that are ranked top 2% are considered as drug-like compounds that should possess the rele-

vant activity. The number of compounds ranked top 2% was reasonable from the practical point of view in this study. The percentage can be varied according to a required enrichment level.

By use of the criteria of drug-likeness described above, the most probable drug-like compounds in the test dataset were clustered.

Enrichment Factor Virtual screening is used to suggest compounds in a dataset, which are ideally of high affinity to a certain receptor. For validation of the suggested compound one has to check the outcome against a sample that is chosen by random. The enrichment factor describes how much higher the proportion of hits (*i.e.* active compounds) is in any given sample of entities compared with the randomly picked sample. When performing virtual screening, the ability of achieving enriched samples is much prized than the ability of predicting single affinities with high accuracy. The formula for the enrichment factor (EF) is usually expressed as follows[8]:

$$EF = V_A/(V(L_A/L)) = LV_A/VL_A$$

Here L and L_A mean the total number of compounds in a library and the number of experimentally verified bioactive compounds in the library, respectively. V and V_A mean the number of compounds predicted as active and the number of known bioactive compounds contained in V, respectively. An enrichment factor of 1 is equally good as random selection.

Software Two programs coded by use of the scientific vector language (SVL) of the software system MOE were made for the present study. The first program DRF is used to analyze the distributions of 2D descriptors and define the ranges of 2D descriptors that cover the sufficient number of compounds. A typical distribution of a descriptor SlogP is shown in Fig. 1. In this figure the range of SlogP that covers 80% of the training dataset is shown. As the distribution curve is a smooth and monomodal one, SlogP is considered as a good descriptor to be used as an identifier of drug-likeness. The second program DRFF calculates the drug probabilities of compounds in the test dataset by use of the ranges of descriptors that were determined by DRF. The source codes of these two programs written by SVL are available from Ryoka Systems Inc.[9] free of charge.

Results and Discussion

The drug-likeness trained by the training dataset was applied in the prediction of active molecules in the test dataset to evaluate the predictive power of this procedure. The summary of the results are given in Table 3. The number of anxiolytic compounds in the test dataset is 274. The trait of anxiolytic compounds was calculated based on the anxiolytic compounds contained in DCUJ. The drug probability as an anxiolytic agent was calculated using this trait for each compound in the test dataset. The top 2% of compounds with the highest drug probability was selected as active compounds (TR). The number of top 2% slightly varies depending on the values of drug probability. In the case of anxiolytic compounds, it was 198. Out of 198 compounds, the number of compounds with actual anxiolytic activity (TR(active)) was 47. Therefore the enrichment factor of anxiolytic compounds based on trait (EF(TR)) was 6.8. The molecular fingerprints of anxiolytic compounds were trained by use of DCUJ. The compounds with the similarity metrics greater than 80% were screened from the test dataset and considered as hits. Out of 123 hits (FP), 46 compounds were experimentally verified active compounds (FP(active)). The enrichment factor attained by use of molecular fingerprint (EF(FP)) was 10.6. The union (FP∪TR) and product (FP∩TR) of the screened compounds by trait and molecular fingerprint show complementarities of these two filters. The number of compounds selected by the product of FP and TR (FP∩TR) is 34 and the number of known active compounds (FP∩TR(active)) contained in it was 26. Although the number of active compounds screened is just 9.0% of the total active compounds, the rate of false positive is quite low and an enrichment factor (EF(FP∩TR)) of 21.8 was attained. If the union of the compounds screened by both of the filters is considered (FP∪TR), the number of hits would increase (67) at the

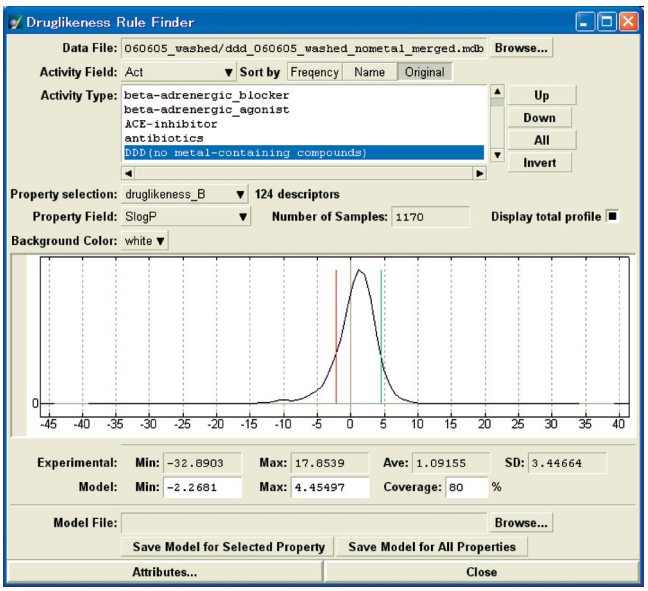

Fig. 1. The Distribution of the SlogP Values of the Drugs Clinically Used in Japan Now

Under the peak between the two vertical lines, 80% of the drugs are included. The horizontal and vertical axes denote the SlogP value and the population of the drugs, respectively.

Table 2. Traits of Drugs Expressed by the Ranges of 14 Typical Molecular Descriptors

	Anxiolytic agent		β-Adrenergic blocker		ACE-Inhibitor		Antibiotics		Drugs[a)]	
Weight	301	377	249	373	368	439	323	751	194	566
SlogP	1.27	4.37	−0.09	1.86	−0.93	0.47	−6.45	1.74	−2.27	4.45
SMR	8.20	9.58	7.26	9.78	9.27	11.63	7.09	18.54	5.03	14.69
TPSA	33	62	46	95	103	107	95	289	26	171
Density	0.73	0.85	0.67	0.75	0.72	0.76	0.74	0.94	0.66	0.89
vdw_area	269	326	299	395	389	454	259	723	206	554
vdw_vol	371	466	368	500	507	604	325	918	253	738
a_acc	2	3	2	4	2	3	2	11	1	6
a_don	0	1	1	2	0	0	1	6	0	3
a_hyd	13	18	12	17	16	22	8	31	7	26
KierA1	10.98	15.03	13.33	19.20	18.71	21.21	12.78	36.33	9.06	27.16
KierA2	4.63	6.50	6.15	9.61	9.19	10.70	5.38	18.18	3.63	12.46
KierA3	2.23	3.12	4.14	7.43	4.80	6.32	2.20	10.96	1.78	7.79
KierFlex	2.61	4.05	4.22	7.28	6.19	7.01	3.11	14.03	2.21	9.12

a) All drug clinically used in Japan.

Table 3. Enrichment Factors Attained for Four Drug Categories by the Drug-Like Profile Analysis

	Anxiolytic agent	β-Adrenergic blocker	ACE Inhibitor	Antibiotics
The number of active compounds in the training dataset	22	31	12	129
The number of active compounds in the test dataset	289	89	27	627
FP[a)]	123	151	55	465
FP(active)[b)]	46	72	14	271
EF(FP)	**10.6**	**44.1**	**77.6**	**7.6**
TR[c)]	198	211	215	201
TR(active)[d)]	47	53	20	115
EF(TR)	**6.8**	**23.2**	**28.3**	**7.5**
FP∪TR	287	290	238	560
FP∪TR(active)	67	74	20	289
EF(FP∪TR)	**6.6**	**23.6**	**25.6**	**6.8**
FP∩TR	34	72	32	106
FP∩TR(active)	26	51	14	97
EF(FP∩TR)	**21.8**	**65.5**	**133.3**	**12.0**

The enrichment factors (EFs) are shown in bold letters. *a*) The number of compounds predicted as active by molecular fingerprint. *b*) The number of known pharmacologically active compounds. *c*) The number of compounds predicted as active by trait. *d*) The number of known pharmacologically active compounds.

expense of low enrichment rate (EF(FP∪TR)) of 6.6. In this case almost 23% of the active compounds in the test dataset were covered. It is assumed that there are several reasons for the discrepancy between L_A and FP∩TR. The compounds in the training dataset are drugs clinically used now. The compounds in the test dataset, however, are compounds that show the pertinent activity but they are not used due to some reasons. This should be a main reason of the discrepancy. That is to say several highly specific molecules are used clinically.

In the case of β-adrenergic blocker, the number of active compounds in the test dataset is 89. The number of the top 2% of compounds with the highest drug probability is 211 of which 53 compounds are known as β-adrenergic blocker. The enrichment rate attained by use of the trait is 23.2. On the other hand, the number of compounds screened by use of molecular fingerprints was 151. The screened compounds contained 72 active compounds. The enrichment rate marked 44.1 in this case. If we get the union of the results by the two filters, 290 compounds were obtained and of which 74 compounds were active. The percentage of hit compounds is 26% and the enrichment rate is 23.6. The hit compounds almost cover 83% of the active compounds in the test dataset. The product of FP and TR gave 72 compounds of which 51 are experimentally verified active compounds. The enrichment factor reached 65.5 and about 57% of active compounds in the test dataset were covered. In this case, the current method showed a good performance.

The results for other two categories of compounds show that the enrichment rates attained by the current method are slightly variable depending on the drug category and the number of compounds in the dataset. In the case of ACE-inhibitor, a high enrichment factor of 133.3 was attained by use of FP∩TR. It covered 52% of the ACE-inhibitors in the test dataset. The enrichment factors for antibiotics, however, were relatively low. Since many different types of compounds that are active against various bacteria were contained in this category, the lower enrichment rate is not surprising.

Conclusions

Virtual screening is a method that complements current advances in high-throughput screening (HTS). It is highly desirable to screen as many compounds as possible in the early stage of drug discovery projects. Even with the modern technology of HTS, the number of compounds that can be handled is limited by several reasons. Virtual screening is undoubtedly becoming a computational counterpart of HTS now. In principle, there is no limit for the number of compounds handled by virtual screening. From the practical point of view the number of the compounds that can be managed by ordinary PC in laboratories could reach at least 10^3 times of the compounds that can be handled by HTS. Although virtual screening is becoming a practical tool in actual drug discovery projects, we have several serious problems to be solved. Using suitable criteria and filters, repeated searches are made through the virtual library with the goal to stepwise decrease its size. Since there is no unique strategy that leads the virtual screening process, there are a number of criteria that have to be considered in virtual screening. Virtual screening is applied mostly in a pragmatic way with the goal to speed up drug discovery in the most economic way. Excluding trivial compounds from the early stage of screening is one of the great benefits of virtual screening. The HTS or ordinary biological screening on especially focused sets of compounds selected by virtual screening can make the entire process of searching and optimizing leads cost-effective and fast. Since virtual screening is an entirely computer-based method, we can employ any sophisticated algorithms. In this paper, however, we used very simple and rather straightforward methods using the molecular fingerprints and trait that are observed in known active compounds. The reliability and enrichment rate could depend on the number of compounds in the training dataset. Therefore if we use a reliable proprietary dataset to train molecular fingerprint and trait that can characterize hit-likeness of compounds, the enrichment rate could increase. The results obtained in this paper, however, unequivocally have shown that the present method could be feasible in the actual drug discovery projects.

Acknowledgements One of the authors (N.H.) is grateful to the Research and Study Program of Tokai University Educational System General Research Organization for financial support.

References

1) "The Merck Index," 13th ed, Merck & Co. Inc., Whitehouse Station, U.S.A., 2001.
2) "Comprehensive Medicinal Chemistry," Elsevier MDL, San Ramon, U.S.A., 2003; http://www.mdli.com/
3) Pharmacast data, Technomics Inc., Tokyo, Japan, 2005.

4) McGregor M. J., Palli P. V., *J. Chem. Inf. Comput. Sci.*, **37**, 443—448 (1997).
5) Labute P., *J. Mol. Graphics. Mod.*, **18**, 464—477 (2000).
6) MOE: Molecular Operating Environment; Chemical Computing Group, Montreal, Canada; http://www.chemcomp.com/
7) Willett P., Winterman V. A., *Quat.Struc.-Act.Relat.*, **5**, 18—25 (1986).
8) Brooijmans N., Kuntz I. D., *Ann. Rev. Biophys. Biomol. Struct.*, **32**, 335—373 (2003).
9) Ryoka Systems Inc., Tokyo, Japan; http://www.rsi.co.jp/kagaku/cs/ccg/dl/index.html