<Review> A Mini-review on Chemoinformatics Approaches for Drug Discovery

Norihito Kawashita\textsuperscript{a,b}, Hiroyuki Yamasaki\textsuperscript{c}, Tomoyuki Miyao\textsuperscript{d}, Kentaro Kawai\textsuperscript{c}, Yoshitake Sakae\textsuperscript{f}, Takeshi Ishikawa\textsuperscript{g}, Kenichi Mori\textsuperscript{h}, Shinya Nakamura\textsuperscript{i}, Hiromasa Kaneko\textsuperscript{d*}

\textsuperscript{a} Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka, 565-0871, Japan

\textsuperscript{b} Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka, 565-0871, Japan

\textsuperscript{c} Laboratory of Pharmaceutical Informatics, Kitasato University School of Pharmacy, 5-9-1 Shirokane, Minato-ku, Tokyo, 108-8641, Japan

\textsuperscript{d} Department of Chemical System Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

\textsuperscript{e} Drug Research Center, Kaken Pharmaceutical Co. Ltd., 14 Shinomiya minamigawara-cho, Yamashina, Kyoto, 607-8042, Japan

\textsuperscript{f} Department of Physics, Graduate School of Science, Nagoya University, Nagoya, Aichi 464-8602, Japan

\textsuperscript{g} Department of Molecular Microbiology and Immunology, Graduate School of Biomedical Sciences, Nagasaki University, 1-12-4 Sakamoto, Nagasaki, 852-8523, Japan

\textsuperscript{h} Medicinal Chemistry Research Labs, Astellas Pharma Inc., Miyukigaoka 21, Tsukuba, Ibaraki 305-8585, Japan

\textsuperscript{i} Department of Pharmaceutical Sciences, Kinki University, 3-4-1 Kowakae, Higashi-osaka, Osaka 577-8502, Japan

\textsuperscript{*} hkaneko@chemsys.t.u-tokyo.ac.jp

Copyright 2015 Chemical Society of Japan
We have reviewed chemoinformatics approaches for drug discovery such as aromatic interactions, aromatic clusters, structure generation, virtual screening, de novo design, evolutionary algorithm, inverse-QSPR/QSAR, Monte Carlo, molecular dynamics, fragment molecular orbital method and matched molecular pair analysis from the viewpoint of young researchers. We intend to introduce various fields of chemoinformatics for non-expert researchers. The structure of this review is given as follows:

1. Introduction,
2. Analysis of Aromatic Interactions,
   2.1 Aromatic Interactions,
   2.2 Aromatic Clusters,
3. Ligand Based Structure Generation,
   3.1 Virtual Screening,
   3.2 De Novo Ligand Design,
   3.3 Combinatorial Explosion,
   3.4 Inverse-QSPR/QSAR,
4. Trends in Chemoinformatics-Based De Novo Drug Design,
5. Conformational Search Method Using Genetic Crossover for Bimolecular Systems,
6. Interaction Analysis using Fragment Molecular Orbital Method for Drug Discovery,
7. Matched Molecular Pair Analysis and SAR Analysis by Fragment Molecular Orbital Method,
8. Chemoinformatics Approach in Pharmaceutical Processes,
9. Conclusion.

Key Words: Chemoinformatics, Drug discovery, Aromatic interactions, Aromatic clusters, Structure generation, Virtual screening, De novo design, Evolutionary algorithm, Inverse-QSPR/QSAR, Monte Carlo, Molecular dynamics, Fragment molecular orbital method, Matched molecular pair analysis

1. Introduction

Structure and properties of compounds are a gold mine of information. Not only physicochemical and chemical graph properties (molecular weight, logP, surface area, the number of hydrogen bond donor and acceptor, graph distance, and others) are calculated from their structures, but also a large amount of information is obtained from dynamic structure changes and from the interaction between compounds and their target proteins. For effective use of a flood of information, obtained from an unlimited number of compounds, information needs to be converted into meaningful data for researchers. Chemoinformatics is a field where information sciences and computational chemistry converge. It includes several research fields and analytical methods such as machine learning, molecular dynamics, molecular orbital method, and statistics [1]-[6]. The information obtained from such process has been used recently for chemical and pharmaceutical researches. Chemoinformatics’ recognition is getting wider and wider since computers have been getting faster and data volume, larger.

In this review, young researchers in Japanese universities and pharmaceutical industries describe reviews on their research fields, focusing on chemoinformatics, and their practical applications to chemical and pharmaceutical researches. The authors of this review are the organizers and the speakers of the 135th annual meeting of the pharmaceutical society of Japan in March 2014. The objective of this review is the introduction of various chemoinformatics related fields for non-expert researchers. This review consists of seven parts as follows:

- The Analysis of Aromatic Interactions,
- Ligand Based Structure Generation,
- Trends in Chemoinformatics-Based De Novo Drug Design,
- Conformational Search Method Using Genetic Crossover for Bimolecular Systems,
- Interaction Analysis using Fragment Molecular Orbital Method for Drug Discovery,
- Matched Molecular Pair Analysis and SAR Analysis by Fragment Molecular Orbital Method,
Chemoinformatics Approach in Pharmaceutical Processes.

2. Analysis of Aromatic Interactions

2.1 Aromatic Interactions

Since intermolecular interactions are key features of stabilization or destabilization in complexes, there are many reports about such interactions. Especially, those with aromatic rings are extensively studied theoretically and experimentally [7]-[10].

For example, a CH-π interaction is recognized as a type of very weak hydrogen-bonding system recently and several studies have showed that the aromatic-aromatic interactions are categorized by ring-ring orientations into stacking interaction, T-shaped interaction and so on (Figure 1) [11][12].

As it is chemically interesting how these interaction categories affect the details of interaction energies, analyses of interactions, such as ab initio molecular orbital calculations, are applied to pairs of aromatic rings in vacuo [13] and practical systems [14][15].

Practically important series of aromatic-aromatic interactions include the interactions between pairs of aromatic residues in proteins. These interactions have been studied extensively and it is found that they are fundamental in various phenomena such as structure stabilities, protein folding, protein-protein recognition, and ligand bindings. For example, Bhattacharyya et al. and Aravinda et al. studied interactions between aromatic residues [16][17]. This knowledge gives us the basis for design and modeling of proteins.

2.2 Aromatic Clusters

In the work of Lanzarotti et al. [19], it was shown that aromatic clusters, aromatic residues close to each other, with three or more aromatic rings are found in many protein structures, expanding pairwise interactions. They, moreover, suggested that these clusters might be important for protein functions, structure stabilities, and ligand bindings.

Yamasaki et al. showed that aromatic clusters are found in pharmaceutically important protein-ligand complexes as well as in apo proteins [20], revealing that these clusters might have a great role for ligand bindings and might help efficient ligand design.

3. Ligand Based Structure Generation

3.1 Virtual Screening

Along with high throughput screening (HTS) assay having been developed, which enables us to screen 100,000 of compounds in one day (or one week) [21], in silico screening methods have also been improved (virtual screening (VS)). These methods have been modified in ways that medicinal chemists can easily apply to their companies’ libraries. At the same time, they can understand philosophies behind those screening methods, trying to modify them toward more sophisticated ones.

For filtering out undesired compounds consistently, initially molecules exhibiting low affinity to target macromolecules should be eliminated. Affinity values are usually estimated with QSAR models because they are easily applied to a vast number of compounds with limited time. Moreover, QSAR utilizes experimental data subsequently. Besides this primary constraint, several criteria such as drug-likeness [22][23], lead-likeness [24], metabolite-likeness [25], and fragment-likeness [26][27] have been applied. Metabolite-likeness aims to select similar compounds to metabolites based on the idea that drugs should be introduced into cells like endogenous metabolites. This criterion, however, is not rule-based. The authors of ref. [25] evaluate similarities between drugs and metabolites with various descriptors. They concluded biochemical pathways should be considered to tighten the search space of chemical space. Fragment-likeness is equivalent to ‘Rule of three’. According to Congreve et al. in [26], for fragments selection from libraries, ‘Rule of three’ states that molecules must satisfy the following constraints: molecular weight <300, number of hydrogen bond acceptors or donors ≤3, ClogP ≤3, number of rotatable bonds ≤3 and topological polar surface area ≤60. Drug-likeness and non-drug-likeness discrimination models based on QSAR have also been studied thoroughly [28]. Machine learning techniques for QSAR/QSPR in drug design was reviewed in a recent publication of Mitchell [29]. In addition to these criteria, chemical structures having toxicophore motifs and
reactive functional groups (electrophiles [30]) to proteins in the human body (always causing promiscuousness [31] of ligands) should be recognized as false positives [32]. False positives seem promising based only on in vitro experiments’ results, but they should not be considered as drugs because they have the undesired features mentioned above. These filters have been implemented with the help of substructure matching techniques. Most of them make use of string representation of molecules, such as SMART and Sybyl/SLN. The filters are adopted in various ways [33][34] for VS. For example, pan assay interference compounds (PAINS) filter, which detects substructures frequently showing promiscuity in HTS, was successfully applied to ligand-based virtual screening for 17β-hydroxysteroid dehydrogenase 2 [35]. In their research, PAINS filter was used to assess whether the proposed inhibitors could interfere with many proteins or not. PAINS filter was used also for constructing high quality libraries of available screening compounds [36].

3.2 De Novo Ligand Design

The idea of Ligand-Based de novo drug design (LBDDD) is to generate (or design) chemical structures satisfying in silico screening criteria (filters) mentioned above. In contrast to structure based de novo drug design (SBDDD), it does not use protein information explicitly. Protein information is indirectly extracted via their ligands’ information. LBDDD usually makes use of existing drugs or molecules, which have high affinity to the target macro molecules [37][38]. These molecules are usually used as reference structures or to extract pharmacophores from them. Regardless of using those reference molecules, generated de novo structures should have preferable pharmacokinetic properties, such as absorption, distribution, metabolism, and excretion. These features can be calculated in silico with QSPR/QSAR models [39]. In short, one of the goals of LBDDD is to design novel molecular structures (if possible, structures having different scaffolds from those of known ligands), whose properties and affinities to macromolecules are plausible in the current hypothesis (i.e., QSPR/QSAR models and filters). However, using weaker hypotheses that do not restrict tightly search space in chemical space, such as rule of 5 [40], always results in generating too many chemical structures (combinatorial explosion), which usually ruins LBDDD.

3.3 Combinatorial Explosion

Combinatorial explosion always causes severe problems with fragment-based ligand design. Fragments are usually functional groups or atoms, depending on the project engaged. Some of fragments used in design may be fixed when these functional groups and the fragments are already known to work for important roles enhancing binding affinity. Urich et al. fixed a building block to design novel kinase inhibitors [41]. Urich et al. extracted building blocks (core fragments) for structure generation from commercial database, which fit the hinge-region in kinases with most likely two or three hydrogen bonding. They ranked the extracted fragments based on structure-based docking simulation. Then, de novo molecules were designed by adding building blocks to each core fragment. The building blocks were selected to fit two hydrophobic pockets of kinases. Although this approach focuses on synthesizability and on structure-based design, fixing one core fragment and connecting building blocks to it seems applicable to ligand-based one. Unfortunately, that strategy (i.e. fixing building blocks, especially one) barely helps to avoid combinatorial explosion. This is because fixing one fragment means that you are able to use one more of it than in the case without fixing a fragment. Using eight fragments without fixing any of them equals fixing one of nine fragments in theory.

The size of chemical space is estimated from $10^{20}$ to $10^{100}$. Calculation of these numbers assume combinatorial features of two dimensional chemical graphs [42]-[44]. There are also some studies considering transformation from the two dimensional structures to three dimensional structures [45]. The developers Ftree-FS define the fragment space as the fragments to be combined and their combination rules. They estimated the size of fragment space derived from a specific database. According to the paper [46], the fragment space of world drug index (WDI) library at that moment was estimated at $10^{18}$ by combining five fragments. They also proposed an efficient algorithm to enumerate similar structures to a reference molecule using dynamic programming, which is equivalent to generating exhaustive structures similar to one used as reference.

When it comes to such exhaustive approaches by combining fragments, well-established generated and collected in a database (GDB) is one of the most famous libraries, (although they used atoms as fragments for compilation of GDB) [47]. GDB17 was compiled in a way that only stable chemical structures consisting of up to 17 heavy atoms (atoms without hydrogen). GDB17 contains 166 billion chemical structures, all of which seem drug-like. Drug-likeness, as well as molecular stability of structures in GDB17, was assured by the generated structures or scaffolds passing through drug-likeness and stability filters [48]. By extrapolating those numbers, Polishchuk et al. [49] estimate the size of drug-like chemical space as $10^{33}$. The algorithm to create chemical graphs in GDB is firstly generating scaffolds using the program GENG [50] developed by Mckay et al., then adding atom-labels and multiple-bonds without violating valency rule to vertices and edges, respectively. By using the same type of generation algorithm, design of novel alkanol amine structures for CO$_2$ absorption was successfully conducted [51]. However, because of
combinatorial explosion, Yamashiro et al. could not use more than 11 heavy atoms. Even when computational speed and resources will have been improved and will be available to us, it is not adequate to overcome combinatorial explosion just by combing fragments.

3.4 Inverse-QSPR/QSAR

One approach to sustain the number of structures to be generated during LBDDD is to combine well-established mathematical graph generation algorithms with a strictly defined narrow region in chemical space [52]. According to Miyao et al., all the structures lying inside the pre-determined region in chemical space are aimed to be enumerated by combining fragments. The method proposed by Miyao et al. can be applied to generate structures in a small region defined by combination of plural regions in chemical space (Figure 2).

Miyao et al. do not consider any synthesizability of generated structures, which is opposite to a lot of structure generators aiming to generate small number of synthesizable molecules [53]. Synthesizability is one of the important concepts in de novo design. In general, synthesizability of de novo structures is assured by either combining existing fragments (building blocks) along known chemical reaction paths [54] or finding retrosynthesis paths from the generated structures [55] toward existing building blocks. In the latter case, it is also important to estimate the synthesizability of those molecules (i.e. retrosynthesizability [56]). Since a secondary goal of de novo design is to inspire medicinal chemists in an unexpected way by proposing novel chemical structures, it is also reasonable to sacrifice synthesizability of proposed structures in exchange for novel scaffolds. In the paper [57], by using ring systems in a database as fragments, around 400,000 structures with high predicted affinity to alpha 2A adrenergic receptor were generated. Some of them shown on the paper have low Tanimoto similarities when compared to the compounds that have desired affinity in the database for model construction.

Since filters with QSPR/QSARs are embedded in VS flow, it looks promising to make use of inverse-QSPR/QSAR methodologies for LBDDD. Inverse-QSPR/QSAR have been studied since 1990 [58][59]. Chemical structures, which have desired properties or activities based on QSPR and/or QSAR models, were proposed. Group contribution method was successfully applied for proposing structures having desired aqueous solubility with partial least square regression [60]. Akutsu et al. solved pre-image problems by using frequency paths as a set of descriptors in kernel space and by using their efficient structure enumerator [61][62]. According to their papers, pre-image problem is a projection problem related to how to retrieve chemical graphs from a desired value in kernel space (Hilbert space defined by kernel function).

Although they do not show examples of practical application with their method, because they focus on proposing efficient algorithms and mathematical flawlessness, the combination between their methods for retrieving chemical graphs and efficient structure enumerator could be applied to inverse-QSPR/QSAR.

The points that determine whether inverse-QSPR/QSAR analysis succeeds or not are: QSPR/QSAR model predictability (1), generation algorithm not to generate duplicated structures as well as to generate exhaustive structures (2), and determination of the region in descriptor space corresponding to the desired objective variables’ values (3).

**Figure 2.** Distinct models defining the desired region in chemical space for drug design. Target region varies from project to project.

Miyao et al. propose to use monotonous changing descriptors (MCDs) [52] for inverse-QSPR/QSAR in order to enhance model predictability (1). Concomitantly, modified canonical construction path method is adopted for exhaustive generation (2). The constraints for structure generation determined via posterior probability density given by a desired objective variable value (e.g. affinity) are calculated based on Bayes’ theorem (3).

MCDs are a set of chemical descriptors whose values monotonously change by adding a fragment during structure generation process, such as molecular weight, the number of a functional group. Using this rather weakly restricted set of descriptors, predictability of QSAR models is expected to be high. They also confirmed that the predictability of aqueous solubility model with MCDs is not bad compared against two dimensional topological descriptors (not shown here). Applicability domain [63] is considered by using ring systems [64] in training data as...
well as by a prior distribution of training data. They showed an example of their proposed workflow through novel ligands design for alpha 2A adrenergic receptor [57].

LBDD in the early stage of drug development is not a transient technique until computational speed and capacity reach a sufficient level where SBDDD can make use of every knowledge about protein-ligand complex, because of combinatorial explosion problems mentioned in the previous section. LBDD may inspire medicinal chemists in their drug design journey and help them understand chemical structures more deeply.

4. Trends in Cheminformatics-Based De Novo Drug Design

In drug discovery process, many properties such as biological efficacy and ADME properties must be optimized simultaneously. To achieve this, medicinal chemists have to synthesize a wide variety of molecules. They often change side-chains and scaffolds (so-called molecular frameworks) to explore a desirable molecule, which meets criteria as a clinical candidate.

Computer chemistry and in-silico drug design play an important role in the discovery of new drugs. De novo drug design has been an active research area over the past decades and many approaches have been proposed, which utilize protein structures and ligand structures. LEGEND [65], LUDI [66], SPROUT [67], LEA3D [68], LigBuilder [69][70], and SYNOPSIS [53] use protein structures, whereas TOPAS [37], CoG [71], Flux [72][73], and NovoFLAP [74] use the structures of known ligands. Some approaches such as LiGen [75] use both for ligand design. Structure-based approaches use a scoring function typically derived from calculated ligand-protein interactions. On the other hand, many ligand-based approaches use a similarity-based scoring function, which calculates molecular similarity between the designed molecule and the known active molecule.

Figure 3 shows the general scheme of evolutionary algorithms reported in the literature [73][76]. Evolutionary algorithms are based on concepts derived from biological evolution, including reproduction, mutation, crossover, and selection and new molecules are designed by repeated application of evolutionary operations. Mutation is crucial for building molecules and the building methods could roughly be classified into one of two categories: fragment-based and atom-based methods. Generally, fragment-based mutations are better considered to generate drug-like structures. Douguet et al [68] and Kawai et al. [76] have respectively investigated atom-based mutation and both of the result yielded unfavorable structures such as invalid hetero-hetero atom bonds.

Recently, Kawai et al. proposed a fragment-based mutation for generating new drug-like molecules [77]. To design molecules, three types of fragments (ring, linker, and side-chain fragments) were defined as building blocks, and a fragment library was prepared from molecules synthesized in the past. As a case study, two GPCRs were selected for computational experiments in which they tried to design ligands from simple seed fragments using the Tanimoto coefficient as a fitness function. A topological fragment descriptor was used for the fitness calculation. The results showed that the algorithm could be used successfully to design new molecules with structural similarity, scaffold variety, and chemical validity [77].

Mishima et al. evolved chemical structures from initial molecules along a path on a two dimensional map spanned by generative topographic mapping in a high dimensional descriptor space [78]. They adjusted the evaluation function so that candidate molecules lie on the subspace. Compounds represented by coordinates in high dimensional space are usually distributed on a certain subspace because of the relationship between descriptors [79][80]. That is why dimensionality reduction methods, such as principal components analysis, are useful for visualization of data in high dimensionality space. Their approach makes sense in order not to generate unwanted structures.

Vinkers et al. reported a de novo design program called SYNOPSIS, which consider synthetic route in designing molecules. The result showed that 10 out of 18 synthesized compounds exhibited cellular inhibitory activity against HIV [53]. IJzerman and co-workers developed a multi-objective evolutionary algorithm for the design and discovering of human adenosine receptor ligands. A pharmacophore model and support vector

![Figure 3. General scheme of evolutionary approaches](image-url)
machine models were used for scoring molecules; successfully obtaining novel ligand with sub-micromolar affinity against adenosine receptor [81]. Recently, Schneider and co-workers used DOGS software to computationally design new molecules for the discovery of novel vascular endothelial growth factor receptor-2 (VEGFR-2) inhibitors. AMG-706, a known VEGFR-2 inhibitor, was used as a design template to obtain a new potent inhibitor with an IC_{50} value of 64 nM [82]. Aforementioned examples show that de novo design, chemical synthesis, and biological evaluation could successfully be integrated into drug discovery projects.

5. Conformational Search Method Using Genetic Crossover for Bimolecular Systems

Molecular simulations have been used in biomolecules such as proteins, DNA, etc. in order to understand their functions. In addition, this approach can be widely applicable to enzymatic chemical reactions and molecular motors. Here, the focus is an all-atom model, based on classical mechanics, for protein systems. Usually, Monte Carlo (MC) and molecular dynamics (MD) are used for the simulations. However, estimation of the stable state by conformational search is difficult, because the energy landscape of the system is characterized by many local minima separated by high-energy barriers. Therefore, conventional MC and MD simulations tend to get trapped in states of local minima. To overcome this difficulty, various sampling and optimization methods for conformations of biomolecules have been proposed, such as generalized-ensemble algorithms (for reviews, see, e.g., [83][84]). Recently, a conformational search method using genetic algorithms has been proposed as well [85]-[87]. In this method, M initial conformations of the system are prepared, where M is the total number of "individuals" in the genetic algorithm, usually considered as an even integer. The following two steps are then alternately performed:

**Step 1.** For the M individuals, regular canonical MC or MD simulations, at a fixed temperature T, are carried out simultaneously and independently for a particular MC or MD step.

**Step 2.** M/2 pairs of conformations are selected from a "parental" group randomly, where crossover and selection operations are performed. The parental group means the latest conformations obtained in Step 1.

While Step 1 is usually based on local updates of conformations, global updates are introduced in Step 2 by genetic crossover. The latter greatly enhances the conformational sampling.

In the stage in Step 2, the selection operation is performed. A superior "chromosome" (conformation) is selected from the parent-child pair. For this selection operation, we can also employ various criteria. In this study, Metropolis criterion [88] is employed, which selects the new child conformation from the parent with the following probability:

\[
w(p \rightarrow c) = \min \left(1, \exp \left( -\beta \left[ E_c - E_p \right] \right) \right)
\]

Here, E_p and E_c stand for the potential energy of the parental and child conformation, respectively, for the parent-child pair. \( \beta \) is the inverse temperature, which is defined by \( \beta = 1/k_B T \) (k_B is Boltzmann constant).

The present method was applied to a protein, which is the helical fragment B of protein A from Staphylococcus aureus (in this paper, we call it protein A) [89]. Although the whole protein A has 60 amino acids, the truncated 46 amino-acid sequences from Gln10 to Ala55 were used in order to examine the helical region, except the terminal region of protein A. For this simulation, the AMBER12 program package was used and the low-point genetic crossover procedure was incorporated. In this simulation, we used 32 individuals (M = 32). For each individual, we performed MD simulations for 90.0 ns (which consisted of 45,000,000 MD steps), and performed the crossover operations 90 times during the simulation.

A similar conformation to the experimental native structure (PDB ID: 1BDD) was obtained, and its root-mean-square distance (RMSD) (only for the backbone atoms) from the native structure was 1.7 Å (see Figure 4). The PDB structure of protein A has three helix regions in the amino-acid sequence of 2-9, 17-29, and 34-46. These helix regions roughly correspond to the high probability regions obtained from the present simulation, and the C-terminal region (Helix III) has the highest helix fraction among the three regions. The stabilities for the three helix structures in protein A have been examined and debated by many researchers. From the simulation results, it is considered that Helix III is the most stable, where the central helix is unstable in comparison with the other two helix structures.

The present method is particularly suitable for highly parallel computing. In the future, this method is going to be applied to larger protein systems.
6. Interaction Analysis using Fragment Molecular Orbital Method for Drug Discovery

With the growth of computer technology and development of new efficient methods, ab initio quantum chemical calculations have been applied to large molecules including proteins. The fragment molecular orbital (FMO) method [90]-[94] is one of the most promising approaches for such large scale quantum chemical calculations. In the FMO method, a target molecule is divided into small fragments, and various physical quantities can be evaluated from calculations of fragment (monomer) and pair of fragments (dimer). For example, each amino acid residue is treated as a single fragment for a typical protein. The total energy is evaluated using the following equation:

$$ E_{\text{total}} = \sum_{I}^{N} E_{I} - (N-2) \sum_{I,J}^{N} E_{IJ}. \quad (2) $$

where $N$ is the number of fragments, $E_{I}$ is the monomer energy of fragment $I$, and $E_{IJ}$ is the dimer energy of fragments $I$ and $J$. This FMO scheme has been extended to various quantum chemical theories, e.g., Møller–Plesset perturbation theory (MP2) [95]-[100], including resolution of the identity (RI) approximation [101]-[104], coupled-cluster theory [105],[106], multi-configurational self-consistent field (MCSCF) theory [107], configuration interaction single (CIS) [108] and its perturbative double (CIS(D)) [109], density functional theory (DFT) [110]-[115], and density-functional tight-binding (DFTB) method [116].

The total energy of Eq. (2) is rewritten by the following equation:

$$ E_{\text{total}} = \sum_{I}^{N} E_{I} + \sum_{I>J}^{N} \Delta E_{IJ}. \quad (3) $$

where $E_{I}$ is the internal energy of fragment $I$, and $\Delta E_{IJ}$ is the inter-fragment interaction energy (IFIE) between fragments $I$ and $J$ [117]. This IFIE can provide very detailed information about the interaction in biomolecules, resulting that many analytical methods based on IFIE have been proposed [118]-[127]. Amari et al. [120] developed a visualized cluster analysis of protein–ligand interaction (VISCANA) using the IFIE. Their idea leads to a systematic analysis of the interactions between a specific protein and various ligands. Mochizuki et al. [121] developed a modified version of configuration analysis for fragment interaction (CAFI), which provides the visualized information of charge transfer in hydrogen bonding interaction. Fedorov et al. [122],[123] introduced an energy decomposition method to FMO scheme, i.e., pair interaction energy decomposition analysis (PIEDA). By this method, the IFIE can be divided into several energy components like electrostatic interaction, charge-transfer interaction, and so on. Kurisaki et al. [124] proposed a two-dimensional matrix expression of the IFIE, by which complicated intramolecular interactions of proteins can be visualized. Tanaka et al. [125] developed a method for statistical correction of the IFIE using classical-mechanical many-body theories. By this method, the screening effects of electro static interaction by amino acid residues and surrounding solvent molecules can be effectively included to the IFIE.

Recently, an analytical method for dispersion interaction in biomolecules [119],[120] was developed by combining the local MP2 (LMP2) method [128]-[131] with the FMO scheme. The name of this method is “fragment interaction analysis based on Local MP2” (FILE). In this method, IFIE is calculated as the sum of two values using the following equation:

$$ \Delta E_{IJ} = \Delta E_{IJ}^{HF} + \Delta E_{IJ}^{LMP2} \quad (4) $$

where $\Delta E_{IJ}^{HF}$ is the interaction energy obtained from the Hartree–Fock (HF) method, which mainly includes electrostatic interaction and hydrogen bonding interaction, and $\Delta E_{IJ}^{LMP2}$ is the interaction energy obtained from the LMP2 method, which mainly includes dispersion interactions or van der Waals interactions. The dispersion interaction of the LMP2 is calculated by summation of pair correlation energies $(\epsilon_{ij})$ between two localized molecular orbitals, so that the IFIE can be transformed to the following equation:

$$ \Delta E_{IJ} = \Delta E_{IJ}^{HF} + \sum_{i \neq j} \sum_{\text{rel}} \epsilon_{ij} \quad (5) $$
where \( i \) and \( j \) are the index of the localized molecular orbital in fragments \( I \) and \( J \), respectively. As a result, the dispersion interaction between two fragments can be decomposed into the contribution of each localized molecular orbital. Furthermore, by applying some type of population analysis to these localized molecular orbitals, it is possible to obtain the dispersion contribution of each atom in the molecule. An illustrative example to a complex of human immunodeficiency virus type 1 protease and a small molecule was given in the previous paper [119]. This method has also been used for interaction analysis of retinoid X receptor which is a member of the nuclear receptor superfamily [132]. As shown here, our FILM can provide much detailed information about dispersion interaction in biomolecules. In particular, this method is very useful for the molecular optimization process of seed compounds in general drug discovery.

7. Matched Molecular Pair Analysis and SAR Analysis by Fragment Molecular Orbital Method

Matched molecular pairs (MMPs) of molecules differ in one part from their chemical structures. Matched molecular pair analysis (MMPA) compiles differences in the arbitrary molecular data of MMPs with identical chemical replacement, such as fluoride to chlorine, from chemical databases to estimate the statistical effect on the molecular data. MMPA has gained increasing interest in the field of chemoinformatics [133]. While MMPA was originally intended to search for frequent chemical replacements from corporate chemical databases [134], a more attractive application is the extraction of medicinal chemical knowledge regarding the effects of specific chemical transformations on physicochemical properties, derived from the real examples of the previous studies [133][135][136].

Interest is also growing in predicting molecular properties based on MMPs [137][138]. Given that increases in data volume increase the power of MMPA, the majority of publications on MMPA are from pharmaceutical companies possessing huge proprietary chemical databases [133][134][139]-[145]. In contrast, relatively few publications on MMPA are from academia [138][146]-[148], likely due to academics’ limited access to large chemical databases. However, the advent of databases such as ChEMBL [149] has increased the availability of chemical structures and experimental values of biological assays, thereby enabling academia to partake in MMPA research. For more information on MMP, please see the detailed review by Dossetter et al. [136], the report of an algorithm for the efficient detection of MMP by Hussain and Rea [150] and the report of a statistical framework for MMPA by Kramer et al. [151].

Further, MMPA was also applied to the biological activities exerted by compounds on their target proteins [137][138]. In addition to MMPA based on two-dimensional chemical structures (2D-MMP), MMPA incorporating three-dimensional information of proteins and ligands (3D-MMP) has also emerged [152]-[155]. The 3D-MMP method identifies the favorable positions for pharmacophores and combines the analyses of structural and activity data to suggest the idea for the synthesis of novel compounds. However, there remain a lot of factors that can be incorporated into 3D-MMP analysis in order to understand the differences in the biological activities of compounds more precisely. For example, the entropic factors associated with the dehydration and the conformational restraints in the ligand binding process contribute remarkably to the binding free energy, and the enthalpic factors attributed to the non-classical interaction and the depolarization effect make a contribution to the interaction energy of the ligand. Nevertheless, these factors are not considered in the current 3D-MMP analysis. The inclusion of these factors might increase the power of 3D-MMP.

The combination of 3D-MMP with quantum chemical calculation of MMPs in the target protein might provide a powerful framework for the analysis of structure-activity relationships (SARs) and deduction of chemical factors at the molecular orbital level. The fragment molecular orbital (FMO) method [91] is a promising solution for the integration of 3D-MMP and molecular orbital methods. The FMO method can calculate, in practical time, the interaction energy between a protein and a ligand via quantum chemical calculation, and even incorporate the whole protein into the calculation. For details of the FMO method, see other sections of this review as well as the comprehensive review by Fedorov et al. [94]. Application of FMO calculation followed by pair interaction energy decomposition analysis (PIEDA) [122] and multivariate statistical analysis to the MMPs for a specific protein enable SAR analysis at the molecular orbital level (SAR-by-FMO) [156]-[158]. The advantage of the FMO method over classical descriptor-based analysis is the incorporation of non-classical interactions and depolarization effects into the SAR analysis [156]. A robust framework of SAR-by-FMO [158] for physical and artificial errors of experimental values and protein-ligand complex models might be effective given the uncertainty in chemical database values, as systematically investigated by Kramer et al. [159].

Both MMPA and SAR-by-FMO can be viewed as a type of local SAR analyses. These methods are suitable to analyze and explain similar chemical modifications, but have difficulties in extending the interpretation to the unknown unique chemical modifications. In general, one of the most difficult aspects of drug discovery process is how we explore unknown chemical space efficiently. Therefore, the next step to advance methods to be more
powerful in the drug discovery process could be how we transfer the knowledge on the chemical modification, accumulated locally in one project, into others. Focusing on the biological activity of the compound to the target molecules, the issue is our deficient knowledge on the ligand binding process. In this point, it is important for us to integrate data mining methods and quantum chemical methods to complement the knowledge. The combination of these methods enables us more systematically to discover interesting experimental data and to unveil novel interaction pattern between biological macromolecules and small molecules. It can deepen the physical and chemical knowledge about the effects of the chemical transformations on the molecular properties, which can pave the way for the future to understand SAR more precisely.

8. Chemoinformatics Approach in Pharmaceutical Processes

It is important to search for a drug compound efficiently. In addition, tablets whose key ingredient is the drug compound must be produced, meeting rigorous quality requirements efficiently and stably in disturbances such as variance of raw materials and changes of production facilities [160][161]. When a tablet in a batch cannot pass the quality tests after several processes such as mixing, tabulating and coating, or even the final product test, all tablets in the batch are wasted. This takes enormous costs. Therefore, the quality of tablets should be monitored and controlled online, but quality measurements, such as active pharmaceutical ingredient (API) content, take too much time. In addition, it is desired that the quality of not just some tablets, but all tablets could be measured in a batch.

When many kinds of drug tablets are handled, it is important to bring the right tablets to the right places. Misclassification, which means that drug tablets are wrongly classified, is a crucial problem. In addition, false tablets, which are dangerous for health, are sometimes marketed illegally. There exist three types of false drug tablets; drug tablets that do not contain any API, drug tablets containing API that are not marked in the packing and drug tablets that contain the marked API but are produced by a different manufacturer [162]. Before acquisition of drug tablets, all tablets must be checked accurately, non-destructively and rapidly. Since false drug tablets production has become high quality nowadays, it is difficult to detect them [163].

Process analytical technology (PAT) [164]-[167] is an important technique for monitoring, developing, controlling and designing critical product quality in the pharmaceutical industry. Near infrared (NIR) spectroscopy [168][169], Raman spectroscopy [170] and so on have been focused to monitor product quality non-destructively in real time. Often, quantitative measurements are desired/required and a calibrated model will have to be developed and implemented [171]-[175]. A regression model is constructed between the quality and the intensity of NIR spectroscopy, and then quality can be predicted using that model. This is called a “soft sensor” [176]-[178]. Soft sensors can be applied not only to pharmaceutical processes, but also to other ones such as chemical processes [63][179]-[181], biological processes [182]-[185] and agricultural processes [186]-[189]. The use of soft sensors and their application to process control are expanding now.

9. Conclusion

In this paper, we reviewed the latest research and methodology of chemoinformatics, which lead to efficient and rational drug design.

To utilize chemoinformatics to drug discovery, the amount and quality of the compound-target interaction are important. As a quality improvement for intermolecular interaction network between ligands and their target proteins, aromatic clusters consisting of CH-π interaction and non-classical interaction obtained from fragment molecular orbital methods has been introduced. The hybrid method of genetic algorithm and molecular simulation for exhaustive stable conformation search for biopolymers has also been described. These improvements affect the calculation of algorithms and fusion of molecular approach and informatics, leading to massive and accurate simulations beyond the old dated computational costs limitation.

Sophisticated information is efficiently applied to make drug discovery faster by chemoinformatics-based De Novo drug design and inverse QSAR/QSPR. Fast and efficient generation of chemical structure satisfying a particular condition is required to achieve such purpose. Moreover, the practical uses and computational tools needed to apply evolutionary algorithms to fragment substitution of lead compound and generation of novel chemical entities are presented.

Chemoinformatics, which unites chemical information on drug compounds and biological information on drug target, would also be applied to dosage design of candidate compounds, proper use of medicines and discovery of new chemical entities. Fusion with chemoinformatics and living body system information obtained by metabolomics [190] and glycomics [191], which is rapidly developing, would make us expect tailor-made or personalized medicine [192] in the future.

These latest methods and tools in this review will be used in drug discovery as usual. Early approval of new medicines promoted by these chemoinformatics...
approaches is expected.

Finally, the author of each chapter is as follows:
1. Introduction (Norihito Kawashita),
2. Analysis of Aromatic Interactions (Hiroyuki Yamasaki),
3. Ligand Based Structure Generation (Tomoyuki Miyao),
4. Trends in Chemoinformatics-Based De Novo Drug Design (Kentaro Kawai),
5. Conformational Search Method Using Genetic Crossover for Bimolecular Systems (Yoshitake Sakai),
6. Interaction Analysis using Fragment Molecular Orbital Method for Drug Discovery (Takeshi Ishikawa),
7. Matched Molecular Pair Analysis and SAR Analysis by Fragment Molecular Orbital Method (Kenichi Mori),
8. Chemoinformatics Approach in Pharmaceutical Processes (Hiromasa Kaneko),
9. Conclusion (Shinya Nakamura).

References and Notes


