Analysing Molecular Landscapes Using Random Walks and Information Theory

B. V. Y. Lee

Abstract

Search heuristics for the in-silico discovery of drug candidates have recently received increased attention. Most of these heuristics such as Simulated Annealing and Evolutionary Algorithms gradually improve molecules by exploiting that the similarity in the molecular structure often relates to the similarity in the properties of molecules such as activity against a target. However, it often remains unproven whether such continuity assumptions actually hold. Generally speaking, there is a need to better understand and assess the properties of molecular search landscapes in order to design/choose appropriate optimization methods to search these spaces.

The theory of combinatorial landscape analysis aims to provide such analysis tools. However, many of the methods proposed in this field require the complete knowledge of the landscape and thus are inappropriate for analysing the huge search spaces of chemical structures. If the size of the search space forbids enumeration, statistical landscape analysis methods are the only available tool.

Following the approach presented by V. K. Vassilev, in this thesis, we propose to estimate landscape properties from random walks using the variation operator of the search heuristic. Once a search heuristic is built, these random walks can be generated with little extra effort. The precision of the obtained results scales with the number and length of the random walks available. Given the data from random walks the following analysis methods can be used: (1) Correlation Length Analysis which reveals the validity of the continuity assumption; (2) Information Complexity which reveals the structural diversity of the search landscape; (3) Multimodality measures which estimate the frequency of local optima for different neighbourhood radii, and finally (4) Neutrality measures which account for the size distribution of plateaus in the landscape. Each measure indicates difficulties for optimization routines encountered when optimizing the objective function.

We apply random-walk based landscape analysis for four search spaces in the context of de-novo drug design: Firstly, we study the properties of three search spaces induced by the mutation operators of the "Molecule Evoluator$^{\mathrm{TM}}$" developed by Eric-Wubbo Lameijer, with an activity model

as an objective function. These three search spaces are of Oestrogen Receptor, Lipoxygenase Inhibitor, and Neuropeptide Y_2 Receptor, respectively. Secondly, we study the properties of a peptide design problem using the software MOE. Here a ligand that binds tightly to a 14-3-3 isoform is searched for.

Contents

Chapter 1

Introduction

Optimization problems are problems of finding the best solution from all feasible solutions. The optimality of a solution can typically be quantified by one or more fitness functions. In many real-world optimization problems, both fitness function and optimization algorithm need to be carefully chosen. An ideal fitness function correlates closely with the chosen algorithm's goal, and is not computationally expensive; while an ideal algorithm consistently demonstrates the ability to converge to good solutions within a reasonable amount of time regardless of the degree of complexity. In our research, the fitness functions have been fixed. Hence, selecting proper algorithms is our prime concern. In order to obtain efficient optimization algorithms, we use fitness landscapes, a concept first introduced by Wright (1932), to model optimization problems.

A landscape is a mapping from a configuration space to the real numbers. A simple kind of configuration space is a graph composed of a set of vertices, and the edges connecting each vertex to its neighbours. Each vertex is considered as a configuration, and a fitness or cost function is applied to get the real value of each vertex. In Chapter 2, we briefly introduce some basic properties of a landscape, including ruggedness, modality, neutrality correlation, local optima, and basins. The former four are intuitive but fundamental properties of a fitness landscape. They help to estimate the difficulties and feasibility of applying evolutionary or optimizing algorithms to a landscape. We use correlation and information analysis to investigate these characteristics. Information analysis is derived from the idea of Vassilev et al. (2000), who defined three information features of a landscape. These three features are information content, partial information content, and information stability, respectively. The former two notions are used in our research. Correlation, an essential property of a fitness landscape itself, together with information content are important indicators of ruggedness. Partial information content is instead an information measure of modality. Density-basin and neutrality blocks are depicting the neutrality of a landscape. Further description of these analysing methods is given in Chapter 3.

In Chapter 4, we discuss about four molecular landscapes. A molecular landscape is a way to represent the properties (in our research, activities) and mutations of simple molecules and protein sequences, which are configurations in the landscape. The results of the experiments carried out on these molecular landscapes are presented in Chapter 5.

The paper ends with some conclusions and future research lines in the last chapter.

Chapter 2

Landscape

2.1 Definition

A landscape consists of two essential elements, a fitness function by which each configuration is mapped to a numerical value, and a rule that defines the neighbours to each configuration. In formal terms given by Stadler (2002),

Definition 1 (Fitness Landscape) *A landscape is an abstraction of a search space composed of*

1. *A set X of configurations,*

2. *a notion χ of neighbourhood, nearness, distance, or accessibility on X,*

3. *a fitness function $f : X \to \mathbf{R}$.*

2.1.1 Travelling Salesman Problem

A typical example of a combinatorial optimisation problem is the travelling salesman problem (TSP). In TSP, a distribution of cities and the costs of travelling from one to any other city have been given. The salesman must visit each city exactly once and then return to the starting point. The total cost of such a tour is simply the sum of each cost from one city to the next in the tour. The problem is to figure out how he could travel at the lowest

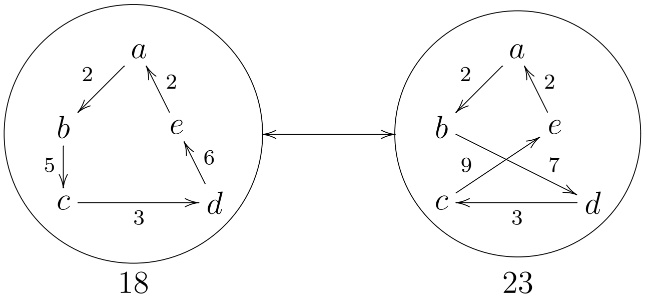


Figure 2.1: Part of a TSP Landscape

total cost. A landscape can be used to model this problem. In this case, all the possible tours are considered as configurations. In particular, they are taken as the vertices of some graph, with the edges connecting them to their neighbours. The neighbours are defined based on certain rules. One of the possibilities to obtain a neighbour of a tour is switching two adjacent cities in this tour; another possibility is taking two edges (A, B) and (C, D) in a tour, and replacing them by the edges (A, D) and (C, B). There are no specific restrictions on defining a rule.

As in Figure 2.1, assume five cities "a, b, c, d, e" and the costs from each one to the others have been given (e.g., from a to b, it costs 2; from b to c, it costs 5). The tour $a \to b \to c \to d \to e$ is one of the configurations(γ_1), i.e., a vertex in the graph. The fitness value of γ_1 is 18. Take the rule ρ_1 that two configurations are neighbours to each other if and only if one tour can be obtained from the other by switching two adjacent cities in the tour. Exempli causa, by switching the order of c and d, we can get the configuration of $a \to b \to d \to c \to e$ (γ_2) with the value of 23, which is one of the neighbours of γ_1 in the graph determined by rule ρ_1. The other edges between vertices are obtained in the same way. It is clear to see the two essential elements of a landscape, as in TSP, a cost function to all the tours, and a rule deciding the neighbours of each tour. Normally, the cost function is fixed when the configurations are specified, while the rule to define a neighbour is not unique. Different rules may therefore result in different landscapes.

2.2 Information Characteristics

Ruggedness, modality, neutrality and correlation are intuitive but fundamental properties of a fitness landscape.

Ruggedness is one of the most important indicators of selecting optimization algorithm. It reflects the relationships between each configuration and its neighbours in the landscape. The greater and more arbitrary fitness variations between adjoining configurations are, the more rugged the landscape will be. The ruggedness of a landscape can be controlled by the threshold set for entropic measures, such as information content.

The modality of a landscape path is another significant feature of a landscape. It shows the moving tendency of the whole landscape, and the tendency to produce local optima. In other words, modality helps to observe the overall picture of the landscape instead of focusing on local details.

In contrast to ruggedness, neutrality represents a connected set of configurations with equal fitness. It is an important indicator of setting the threshold scale for analysing the ruggedness and modality.

Correlation plays an important role in landscape analysis. It indicates the dependence between two configurations in the landscape. In a landscape consisting of an extremely large number of configurations, correlation is a natural approach to estimate the global structure. Correlation measures are further discussed in section 3.1.

2.3 Local optima

In TSP, finding a tour solution that is no worse than all its neighbours is relatively easy. In combinatorial optimisation, a general term referring to such a solution is *local optimum.* The best of all these local optima are accordingly called *global optima.*

Unlike a global optimum, which is concerned with the whole configuration space X, a local optimum is only in regard to a neighbourhood in X. The formal definition adapted from Stadler (2002) is:

Definition 2 (Local minimum) *A configuration $\hat{x} \in X$ is a local minimum if there is a neighbourhood $N(\hat{x})$ such that $f(\hat{x}) \leq f(y)$ for all $y \in N(\hat{x})$.*

Looking for local optima could be an effective way to approach the global optima as every local optimum can be considered as a potential global optimum. While on the other hand, in many cases, for example, doing hill climbing, when the landscape is very “rugged”, the search for global optima will be easily trapped at local ones.

The number of local optima is an indicator for the ruggedness of a landscape. As Palmer (1991) suggested, a landscape X is considered rugged if the number M_X of local optima scales exponentially with some measure of “system size” such as the number of cities in a TSP. It should be noted though that this way of judging a landscape’s ruggedness may lead to confusion sometimes. For instance, a flat landscape, whose local optima include all its elements, will be defined as a rugged one by this means.

The exact number of local optima M_X can be obtained by generating and checking Definition 2 through the whole landscape. The operation of comparing one configuration with all its neighbours needs to be applied to each and every configuration existing in the landscape, i.e.,

$$M_X = \sum_{x \in X} \left[\bigwedge_{y \in N(x)} f(x) \leq f(y) \right] .$$

In this formula, the construct $[\cdot]$ is used to test the condition. If the condition is true, it will return 1, otherwise, it will return 0. In search space X, M_X is the number of configurations x, which satisfy the criterion that the fitness value $f(y)$ of any y belonging to the neighbourhood $N(x)$ of x is no smaller than the fitness value $f(x)$ of x.

In general cases, this can be quite expensive in terms of run time and memory. Thus, in practise, estimating the value of M_X is usually taken as the compromise but effective way. In this way, computations will be applied only to randomly chosen configurations of a landscape. M_X can instead be approximated by,

$$M_X \approx \frac{|X|}{n} \sum_{i=1}^{n} \left[\bigwedge_{y \in N(x_i)} f(x_i) \leq f(y) \right] ,$$

where the x_i are n uniformly randomly chosen configurations.

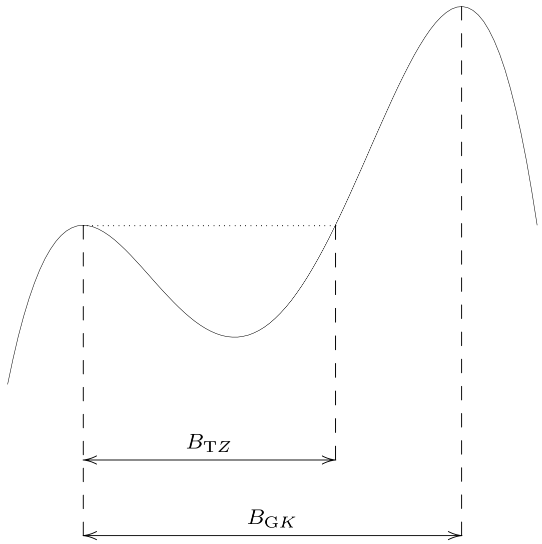


Figure 2.2: Different definitions of Basin

2.4 Basin

Basin is an important notion associated with local optima. So far, there is no definition of basin accepted in full generality. We take the one given by Garnier and Kallel (2002) in our research, which is displayed as B_{GK} shown in Figure 2.2.

Definition 3 (Attraction Basin) *The attraction basin of a local minimum x_j is the set of points $y_1, ..., y_k$ of the search space X such that a steepest descent algorithm starting from y_i for $(1 \leq i \leq k)$ ends at the local minimum x_j. The normalised size of the attraction basin of the local minimum x_j is then equal to $k/|X|$, with $|X|$ standing for the number of elements contained by X.*

Whereas Törn and Žilinskas (1989) suggested B_{TZ} as seen in Figure 2.2, that the basin of x corresponding to a single connected set of the region of attraction is the maximal level set contained in this region of attraction.

As well as the number of local optima, the distribution of basin sizes is also a measure of a landscape's ruggedness. The fewer basins of small sizes, the smoother the landscape will be.

Chapter 3

Information Measures

All the analysis methods in this discussion are based on time series obtained by random walks on a landscape, and each landscape point is a genotype (a string of 0s and 1s).

3.1 Correlation

Correlation indicates the linear relationship between two random variables. This concept is often used in statistics, as well as in landscapes. Correlation measures are by the far the most accessible indicators of ruggedness.

A variety of correlation methods have been applied to landscapes. For instance, Fitness Distance Correlation (FDC) introduced by Jones and Forrest (1995). FDC measures the correlation between the fitness values of a function under investigation and the distance to the optimal solution. Given a set $F = \{f(x_1), f(x_2), ..., f(x_n)\}$ representing the fitness values of the configurations in the search space and a set $D = \{d(x_1), d(x_2), ..., d(x_n)\}$ representing the corresponding distance of those points from the optima solution, the correlation coefficient, ρ is calculated as

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \mu_F)(d(x_i) - \mu_D)}{\sigma_F \sigma_D},$$

where $\mu_F, \mu_D, \sigma_F, \sigma_D$ are the means and standard deviation of F and D, respectively.

In our research, autocorrelation of fitness values obtained by random walks is adopted. Different from FDC, autocorrelation is the cross-correlation of fitness values. A walk correlation is one way to get the autocorrelation. Two sets of variables are required in the computation of the walk correlation. One is randomly chosen from the set of vertices of a landscape, while the other is obtained along a random walk by starting from the former. Let X_s denote the values of the nodes at time s, where the random walks start, and X_t be the values of the process at time t, where t may be an integer for a discrete-time process or a real number for a continuous-time process. The autocorrelation coefficient is defined by time series as:

$$\rho(t, s) = \frac{\mathrm{E}[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s}$$

where "E" is the expected value operator, $\mu_s, \mu_t, \sigma_s, \sigma_t$ are the mean averages and standard deviations at time s and t respectively.

For a considerably large landscape, the value of the walk correlation should start from 1 at $s = 0$, and decline to 0 (normally sufficiently close to 0) as s approaches to infinity. The more rugged the landscape is, the faster the descent will be.

These properties can also be applicable to a small landscape if the transition kernel is ergodic, and $p_{x,x} \neq 0$ for any $x \in X$ ($p_{x,x}$ is the probability of remaining in the same $x \in X$).

Obviously, there is no meaningful definition for correlation for a flat landscape, since the standard deviation will be 0 in a flat landscape.

3.2 Information Analysis

Vassilev et al. (2000) proposed information analysis for studying the structure of fitness landscapes. In this idea, three information characteristics measures are defined, which are termed information content, partial information content, information stability, all of which are threshold-based indicators.

Consider a time series $\{f_t\}_{t=0}^{n}$, which contains the fitness values in real numbers obtained by random walks on the landscape, where f_t is the fitness value of the genotype x_t achieved at step t from the starting point x_s.

	x_0	x_1	x_2	x_3	x_3	x_4	x_5	x_6	x_7	x_8	
	2.5	7.1	3.8	5.7	1.2	4.5	4.3	4.7	8.0	8.1	
$S(0)$		$\nearrow$	$\searrow$	$\nearrow$	$\searrow$	$\nearrow$	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\epsilon = 0$
$S(0.3)$		$\nearrow$	$\searrow$	$\nearrow$	$\searrow$	$\nearrow$	$\rightarrow$	$\nearrow$	$\nearrow$	$\rightarrow$	$\epsilon = 0.3$
$S(3)$		$\nearrow$	$\searrow$	$\rightarrow$	$\searrow$	$\nearrow$	$\rightarrow$	$\rightarrow$	$\nearrow$	$\rightarrow$	$\epsilon = 3$

Figure 3.1:

$\{f_t\}_{t=0}^n$ represents a path in the landscape. For a given parameter ϵ, it can be transformed into the string $S(\epsilon) = s_1 s_2 ... s_n$ of symbols $s_i \in \left\{\bar{1}, 1, 0\right\}$, where

$$s_i = \begin{cases} \bar{1}, & \text{if } f_i - f_{i-1} < -\epsilon \\ 1, & \text{if } f_i - f_{i-1} > \epsilon \\ 0, & \text{otherwise} \end{cases}$$

for $\epsilon \in [0, \ell]$ where ℓ is the maximum difference between two fitness values.

The parameter ϵ is a threshold that determines the accuracy of the calculation of the string $S(\epsilon)$. How ϵ influences upwards/downwards behaviour in the analysis is perspicuously drawn in Figure 3.1.

The fitness variation is considered to be moving upwards if $s_i = \bar{1}$, which means the fitness value of the preceding genotype is smaller than its successor, and the difference is greater than ϵ; if $s_i = 1$, the movement is then considered downwards with a difference greater than ϵ; while otherwise the movement is considered as level.

Therefore in Figure 3.1, when ϵ is set to be 3, it moves upwards from the genotype x_0 to x_1 as $f_1 - f_0 = 4.6 > 3$, while it moves downwards from x_1 to x_2 as $f_2 - f_1 = -3.3 < -3$. And from x_2 to its successor x_3, it is a flat movement, while it is considered upwards by both $\epsilon = 0$ and $\epsilon = 0.3$.

It is easy to see from the Figure 3.1 that, the smaller the value of ϵ is, the more sensitive the movement will be to the differences between the fitness values, and the more precisely the landscape can be depicted, while the peaks along the path will also multiply, as well as the obstacles to approach a good solution.

3.2.1 Information Content

Information content is an entropic measure defined by Vassilev et al. (2000).

$$H(\epsilon) = -\sum_{p \neq q} P_{[pq]} \log_6 P_{[pq]}.$$

Both p and q are elements from set $\left\{\bar{1}, 1, 0\right\}$. The probabilities $P_{[pq]} = \frac{n_{[pq]}}{n}$ are frequencies of the possible blocks pq, where n_{pq} is the number of occurrences of pq in $S(\epsilon)$. The base of the logarithm is taken as 6 since this is the number of all possible combinations of pq, as $p, q \in \left\{\bar{1}, 1, 0\right\}$ and $p \neq q$.

This measure estimates the diversity of local landscape shapes along the landscape path, in other words, it denotes the ruggedness of the landscape path represented by $S(\epsilon)$. The parameter ϵ works as a threshold in terms of zooming in or out from the landscape. By increasing or decreasing ϵ, we will get a closer look or a more general picture.

3.2.2 Partial Information Content

Partial information content is determined by a new string $S'(\epsilon)$ constructed from $S(\epsilon)$ which is associated with the time series $\{f_t\}_{t=0}^{n}$. In $S'(\epsilon)$, all the elements of 0 in $S(\epsilon)$ are removed, and all the contiguous elements with the same value will be taken as one element. For example, the resulting string $S'(\epsilon)$ of an $S(\epsilon) = 1\bar{1}\bar{1}1\bar{1}10110$ is $1\bar{1}1\bar{1}1$. Clearly, $\nu(\epsilon)$ as the length of $S'(\epsilon)$ is the number of extrema along the landscape path, which indicates the modality of the corresponding landscape path.

With $\nu(\epsilon)$, we have the definition of partial information content $M(\epsilon)$ given by Vassilev et al. (2000) as (3.1):

$$M(\epsilon) = \frac{\nu(\epsilon)}{n} \tag{3.1}$$

where n is the length of $S(\epsilon)$.

Thus the partial information content $M(\epsilon)$ is 0 if the landscape path is flat, or monotonously increasing or decreasing, while $M(\epsilon)$ is approaching 1 as the modal variations grow. When the modal variations of the landscape path reaches the maximum, $M(\epsilon)$ is 1.

3.3 Neutrality measures

Two kinds of neutrality measures adopted from Leier and Banzhaf (2003) are discussed here, density-basin information and lengths of neutrality blocks, both of which are based on the string $S(\epsilon)$ described in Section 3.2.

Contrary to information content, density-basin information $h(\epsilon)$ represents the ratio of relatively flat and smooth areas. It is defined in Leier and Banzhaf (2003) as

$$h(\epsilon) = - \sum_{p \in \{\bar{1},0,1\}} P_{[pp]} log_3 P_{[pp]}$$

Same as in Leier and Banzhaf (2003), the lengths of neutrality blocks $P_{[0]}, P_{[00]}, P_{[000]}$ are instead taken as the neutrality measure in our research, where $P_{[0...0]}$ means the frequency of blocks 0...0 in $S(\epsilon)$. It is a more expressive and direct way than density-basin information to picture the neutrality of a landscape path.

Chapter 4

Molecular Landscapes

Fitness landscapes can be used to describe and investigate chemical information. A molecular landscape, which is the visualization of chemical space, makes it easier to comprehend the transition between chemical structures by mutations. Two essential components of a molecular landscape are the representation of molecules and the fitness function. Due to the large population of molecules and unpredictable mutations of a molecule, it is normally too expensive to apply an exhaustive search algorithm to the search space in a molecular landscape. Based on the assertion of Brown (2009) "similar molecules will also tend to exhibit similar properties", which is also known as *similar-structure, similar-property principle*, or simply referred to as the *similar-property principle*, similar (neighbouring) configurations will have similar performance relative to a goal in a molecular landscape. Heuristic algorithms can therefore be good options for solving optimization problems in molecular landscapes.

In our research, we will go through two of the main methods developed in drug design, similarity approaches and ligand-protein docking.

4.1 Activity of Drug Like Molecules

The "similarity approach" is to search molecules with the similar activities to the target molecule, which might replace the target molecule with much lower costs. Lameijer et al. (2005) present the "Molecule Evoluator" as a piece

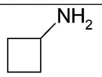
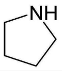
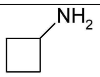
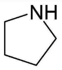
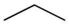
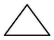
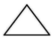
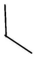
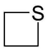
Mutation name	**Initial structure**	**Final structure**	**Initial TreeSMILES**	**Final TreeSMILES**
Add atom			...(C(H)(*H*)(C…	...(C(H)(**N(H)(H)**)(C…
Insert atom			...(C(H)(H)(C...)...	...(C(H)(H)**(N(H)**(C...)**)**...
Delete atom			..(C(H)(*N(H)(H)*)(C...	...(C(H)(**H**)(C...
Uninsert atom			...(C(H)(H)*(N(H)*(C...)*)*...	...(C(H)(H)(C...)...
Increase bond order			...(C*(H)*(H)(C*(H)*(H)(...	...(C(H)(**=**C(H)(…
Create ring			(C(*H*)(H)(H)(C(H)(H)(C(*H*)(H)(H)))	(C(**1**)(H)(H)(C(H)(H)(C(**1**)(H)(H)))
Decrease bond order			...(C(H)(=C(H)(…	...(C**(H)**(H)(C**(H)**(H)(...
Break ring			(C(1)(H)(H)(C(H)(H)(C(1) (H)(H)))	(C(C(H)(H)(H))(H)(H)(C(H)(H)(H)))
Mutate atom			...(C(H)(H)(*C(H)(H)*(C...	…(C(H)(H)(**S**(C...

Figure 4.1: Schematic Overview of the Different Mutations in the Molecule Evolator$^{\mathrm{TM}}$

of software developed for drug design. There are 9 kinds of basic molecular mutations in the Molecule Evoluator, as described in Figure 4.1. The activity value of each molecule is regarded as a configuration in the landscape, while the 9 mutations are taken as the transformations between one molecule and its neighbours. An expert user is used as the fitness function in the landscape.

Kruisselbrink et al. (2008), on the other hand, suggest an automated design of molecules as another option. Instead of depending on the experience and intuition of an expert, a computer programme performs an automated selection on the basis of some preset rules. As compared to Lameijer et al. (2005), two additional mutations are used by Kruisselbrink et al. (2008). These are the "add group" and "delete group" operations as shown in Figure 4.2. This new approach avoids biased decisions and lessens time consumption, both of which may be caused in the interaction in an expert system. However, this design might encounter the difficulties to filter out obviously bad solutions, which are normally not obstacles to an expert. Moreover, it is difficult to include all rules a chemical expert may apply to select a molecule into an automatically evaluated criterion. This approach is thus suggested as an alternative and supplementary method to an expert interaction system, rather than a replacement.

For this study, three test-cases are used, which are the Oestrogen receptor, the Lipoxygenase inhibitor, and the Neuropeptide Y_2 receptor. The goal is to find ligands for these target based on the activity models generated. The activity of ligand molecules was measured for different molecules in the search space, and served as the fitness function.

4.1.1 Oestrogen Receptor

The oestrogen receptor (ER) is a ligand-activated transcription factor that mediates the effects of the steroid hormone 17β-oestradiol in both males and females (Enmark and Gustafsson 1999). It is thought to play a crucial role in the regulation of many life processes, including development, reproduction an normal physiology (Korach et al. 1996).

4.1.2 Lipoxygenase (LOX) Inhibitor

Lipoxygenases (LOX) belong to a heterogeneous family of lipidperoxidizing enzymes and are involved in the biosynthesis of mediators of inflammation. LOX is involved in the metabolism of fatty acids (and thus simply the fats we are eating) (Kruisselbrink et al. 2009).

4.1.3 Neuropeptide Y_2 (NPY2) Receptor

Neuropeptide Y receptors are a class of G-protein coupled receptors which are activated by the closely related peptide hormones neuropeptide Y, peptide YY and pancreatic polypeptide (Michel et al. 1998). These receptors are involved in control of a diverse set of behavioural processes including appetite, circadian rhythm, and anxiety. Specifically, Neuropeptide Y_2 receptor antagonists are involved in obesity/weight control (Kruisselbrink et al. 2009).

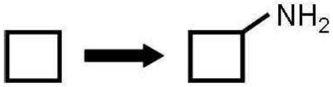
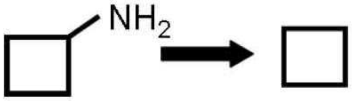
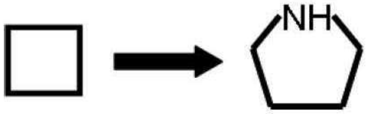
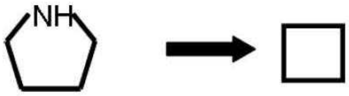
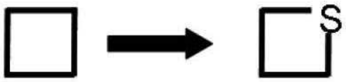
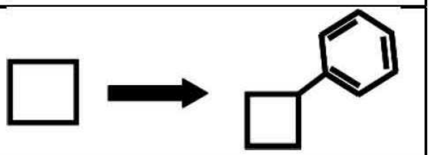
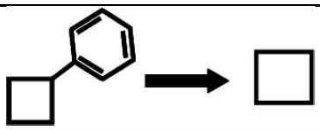
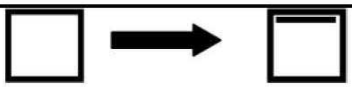
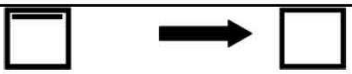
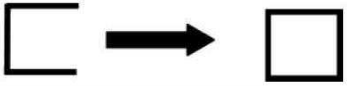
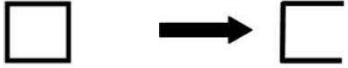
Mutation name	Effect
Add atom	
Remove atom	
Insert atom	
Uninsert atom	
Mutate atom	
Add group	
Remove group	
Increase bond order	
Decrease bond order	
Make ring	
Break ring	

Figure 4.2: Schematic Overview of the Different Mutations in the Molecule Evoluator

4.2 Finding a ligand 14-3-3 γ Isoform Receptor

Ligand-protein docking is an approach to simulate the physical processes involved in a ligand binding to a protein binding site or the prediction of the structure of receptor-ligand complexes (cf.Brown (2009); Brooijmans and Kuntz (2003); Kitchen et al. (2004)).

The target of the second case study is to find a ligand to the 14-3-3 protein. This protein is responsible for cell growth in tumours, and to find a ligand would be interesting in order to cure cancer. In the research project with Drs. H. S. Faddiev (LIACS) and Prof. Herman Spaink (Molecular Cell Biology Department, Leiden University), the MOE software was used in order to find a peptide that binds only to the 14-3-3 γ isoform. The reason we focus on peptides is that the research group of Molecular Cell Biology can synthesize them and test them on cells.

The following description is taken from Faddiev (2008):

> The 14-3-3 γ isoform 4.3 is taken from the crystal structure of the human 14-3-3 bonded to R18 peptide. It is easily observed that atoms and bonds located in the binding grove are highly conserved. This mean that all iso-forms share ligand recognition process inside binding grove and the ligand selectivity is due to the amino acids outside the binding grove. The ligand to be designed should extend from binding grove to reach active surfaces on top and bottom of the grove. ODA (optimal docking areas) outside the binding grove based on atomic disolvetion [sic] parameters were identified in the work of Yang et al [7]. To reach two surfaces located below and above the binding groove a 23-amino acids peptide is placed through the binding grove close to ODA surfaces, as shown on Figure 4.4.

Instead of 23, we have only 5 amino acids in our experiments.

A peptide is encoded as a string of length 5 from the alphabet of amino acids out of the set {ALA, ARG, ASN, ASP, CYS, GLU, GLN, GLY, HID, HIE, HIP, HIS, HYP, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL}. The search space is the space of all possible peptides of a

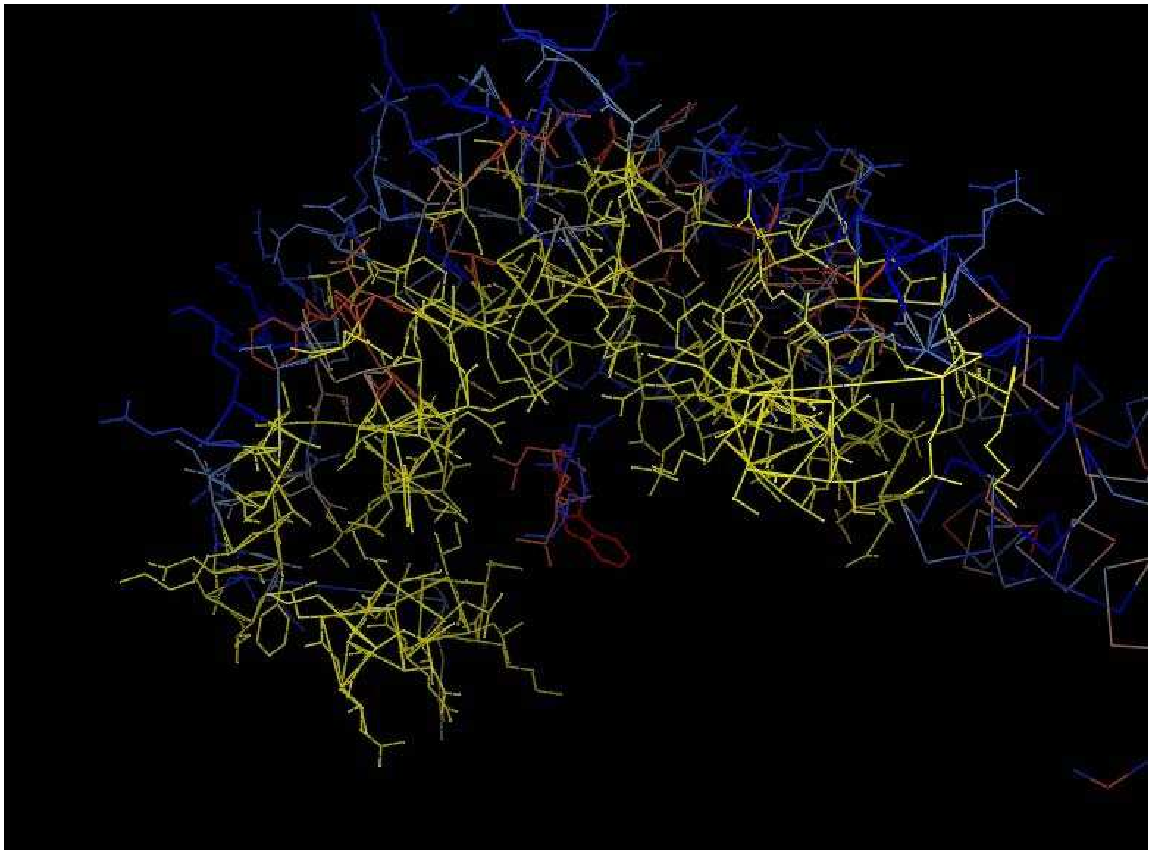


Figure 4.3: Crystal structure of human γ isoform

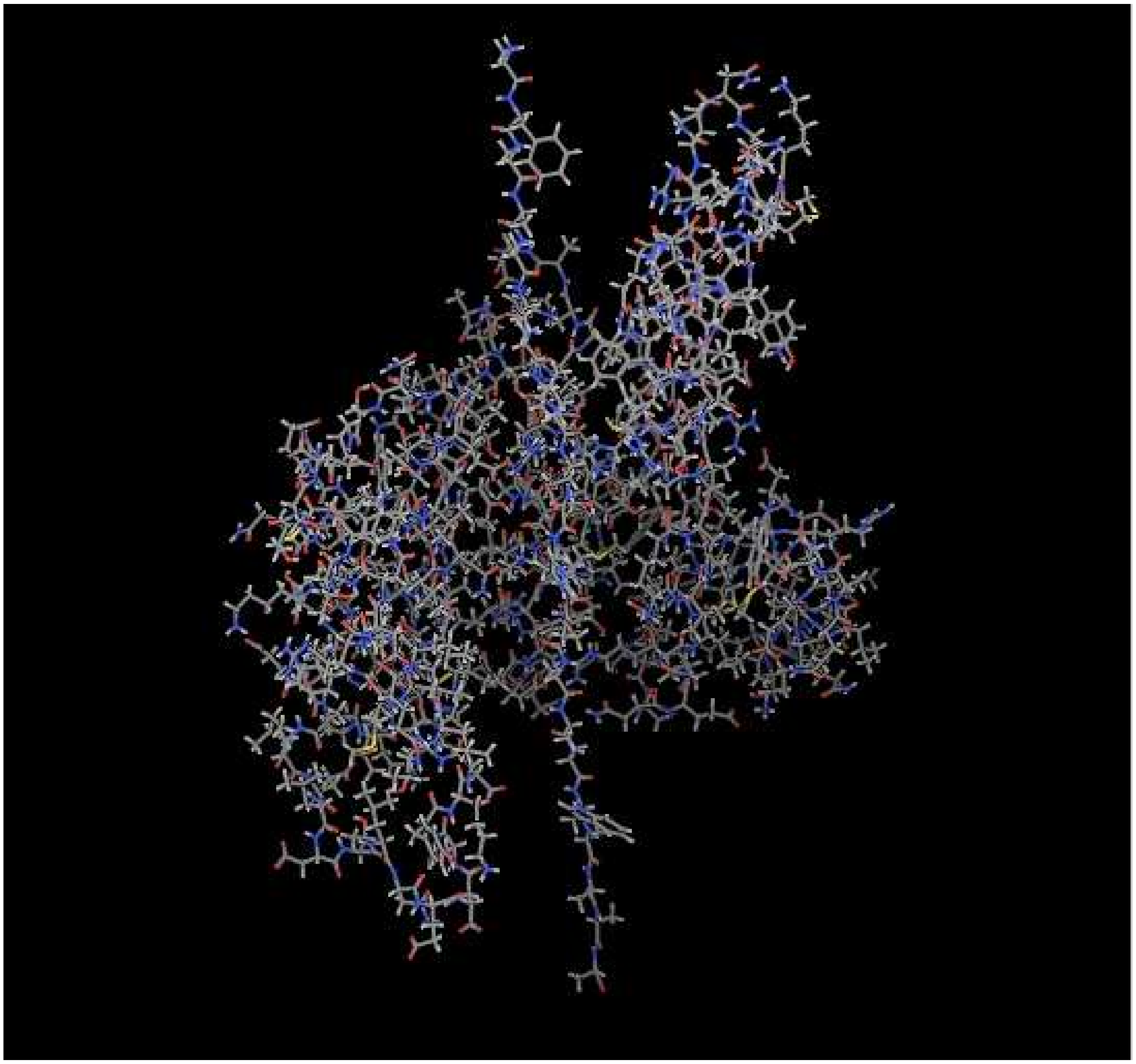


Figure 4.4: The 23 aminoacid peptide is located in near extended conformation inside the binding grove of the 14-3-3 γ isoform

prescribed length. A mutation is to change one amino-acid in the string, e.g.:

$$\text{HIE HIP ILE LEU SER} \rightarrow \text{HIE PRO ILE LEU SER.}$$

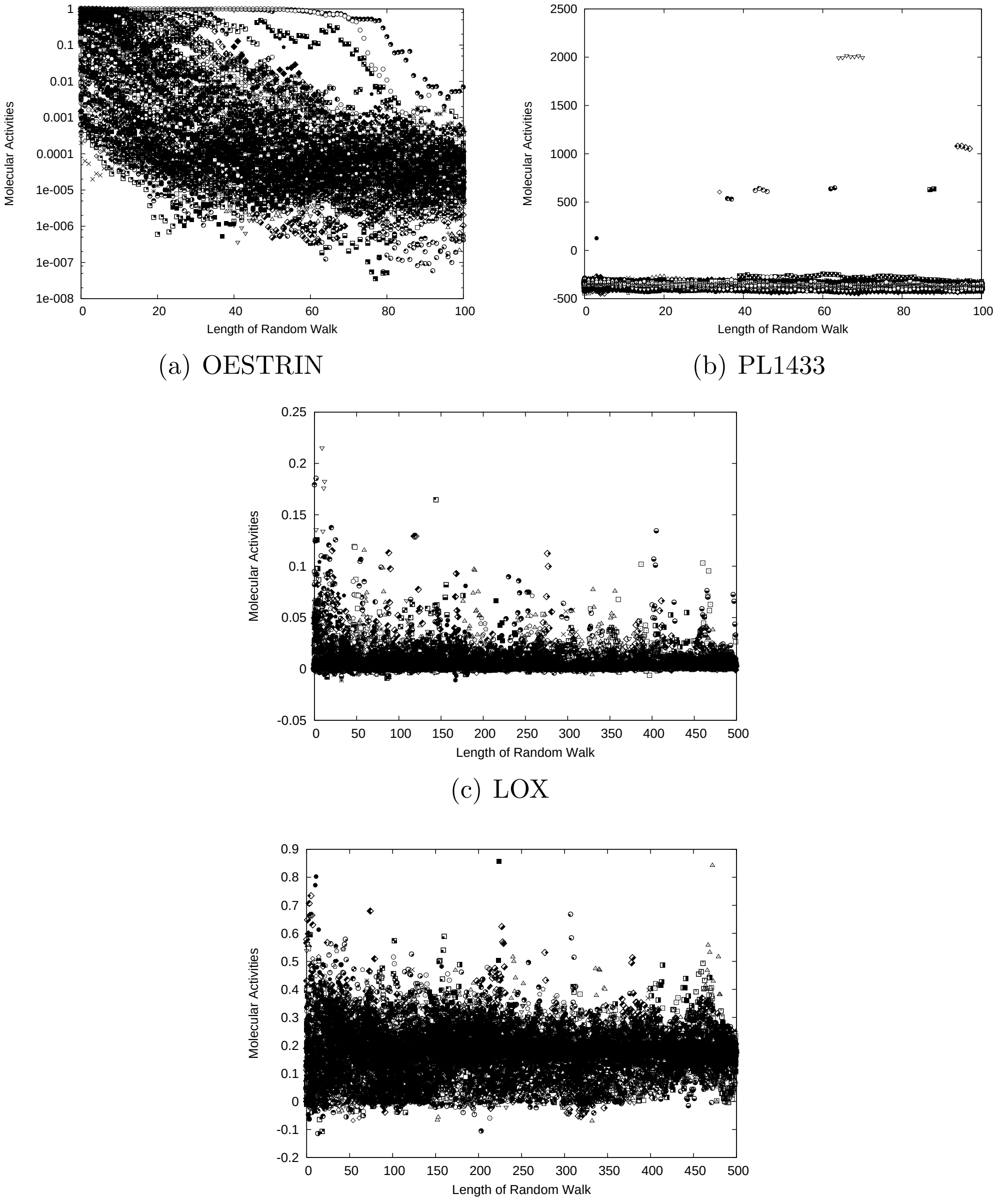
In silico, one docking experiment takes circa 1min.

Chapter 5

Experiments and Results

In Chapter 3, we discussed about several information measures, including correlation, information content, partial information content and neutrality measures. In this chapter, these four measures are applied to the four molecular landscapes mentioned in Chapter 4.

Following the approach of Vassilev et al. (2000), we propose to estimate landscape properties from random walks using the variation operator of the search heuristic. Starting from a randomly chosen configuration, each random walk traces the mutations on the landscape. Along the path of a random walk, each configuration precedes one of its neighbours obtained by a random operator. On Oestrogen Receptor Landscape OESTRIN, and Peptide Receptor Landscape PL1433, we have 100 and 151 walks performed respectively, each of which consists of 100 steps. For both Landscape LOX and Landscape NPY, who represents the activities of inhibitors of Lipoxygenase and ligands of the Neuropeptide Y_2 receptor respectively, we performed 100 walks, each of which consists of 499 steps. Figure 5.1 gives an overview of random walks on these four landscapes. It presents the molecular activities, which are fitness values in the landscapes, along the random walk path. Among the plots, the activities on oestrogen receptor of drug like molecules are depicted in log scale.



(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.1: Fitness Values in Molecular Landscapes

It can be arranged in the matrix (5.1) as follows,

$$\begin{matrix} x_{0,0} & x_{0,1} & x_{0,2} & \dots & x_{0,t} & \dots & x_{0,n} \\ x_{1,0} & x_{1,1} & x_{1,2} & \dots & x_{1,t} & \dots & x_{1,n} \\ x_{2,0} & x_{2,1} & x_{2,2} & \dots & x_{2,t} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_{m,0} & x_{m,1} & x_{m,2} & \dots & x_{m,t} & \dots & x_{m,n} \end{matrix} \tag{5.1}$$

where each row represents a random walk, while each column represents a step in the random walk, e.g, $x_{0,1}$ is the configuration obtained by 1 step from starting point $x_{0,0}$; $x_{0,2}$ is the configuration obtained by 2 steps from starting point $x_{0,0}$, and by 1 step from point $x_{0,1}$; $x_{m,t}$ is the one obtained by t steps from starting point $x_{m,0}$, and by 1 step from point $x_{m,t-1}$.

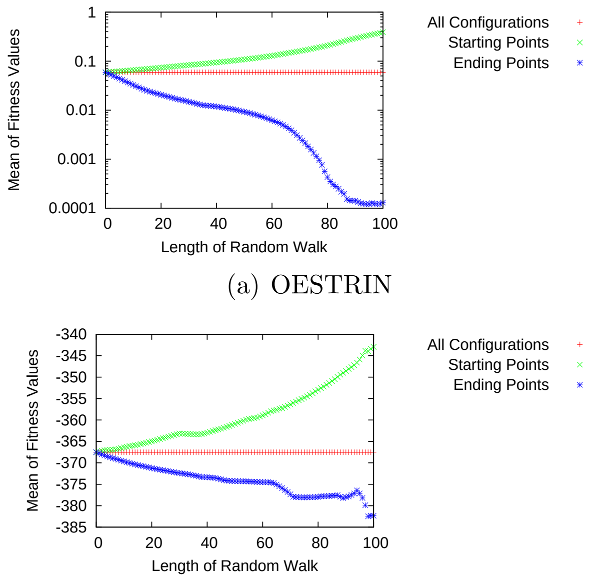
5.1 Correlation

Molecular landscapes are considered to be statistically isotropic. That means in molecular landscapes the variance of statistics does not depend on the starting points chosen but only on the distance between the populations. A sufficiently long unbiased random walk in such a landscape can determine the correlation. Based on this assumption, Weinberger (1990) proposed the autocorrelation function:

$$\rho_s = \frac{\mathrm{E}[f_t f_{t+s}] - \mathrm{E}[f_t]\,\mathrm{E}[f_{t+s}]}{\mathrm{V}[f_t]}\ ,$$

where $\mathrm{E}[f_t]$ and $\mathrm{V}[f_t]$ are the expectation and the variance, respectively, of the time series. The autocorrelation function indicates the correlation between points that are separated by a distance s.

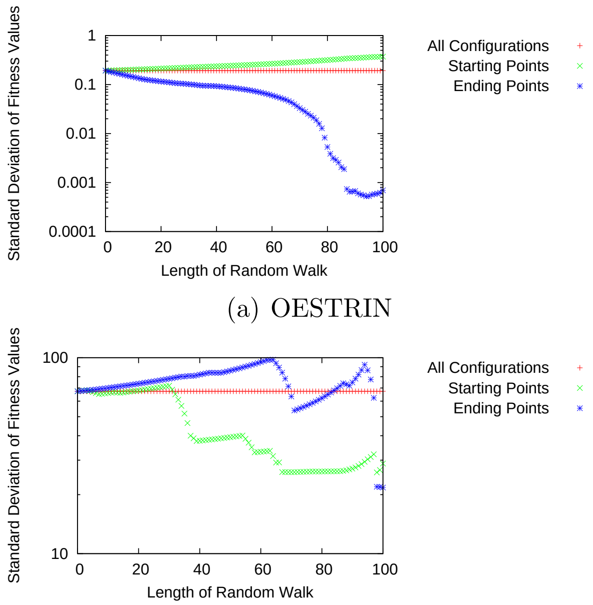
Due to the large number of configurations contained in a molecular landscape, the quantities of time series we gained are relatively small. This may result in time-varying volatility in time series. In other words, the mean values and standard deviations are time varying instead of being stationary. This can be seen in plot 5.2 and 5.3, which show the relationship of means and of standard deviations among starting points, ending points and all configurations in random walks on OESTRIN and Landscape PL1443. Same as in the plots 5.1, we use log scale for OESTRIN.



(a) OESTRIN

(b) PL1443

Figure 5.2: Mean Fitness Values



(a) OESTRIN

(b) PL1443

Figure 5.3: Standard Deviations of Fitness Values

We therefore use the function (5.2) for our analysis:

$$\rho_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\mu_0)(T^s f(x_i)-\mu_s)}{\sigma_0\sigma_s}, \tag{5.2}$$

where $f(x_i)$ is the fitness value of x_i on the landscape X, $T^s f(x_i)$ is the fitness value obtained after s steps of a random walk starting from x_i, and $\mu_0, \mu_s, \sigma_0, \sigma_s$ are the mean averages and standard deviations at T^0 and T^s, respectively.

The mean values in our analysis are calculated by (5.3):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f(x_i), \tag{5.3}$$

while the deviation is denoted by (5.4):

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\mu)^2}. \tag{5.4}$$

According to (5.1), the function can be written more specifically as (5.5):

$$\rho_s = \frac{\frac{1}{n+1}\sum_{i=0}^{n}(x_{i0}-\mu_0)(x_{is}-\mu_s)}{\sigma_0\sigma_s}, \tag{5.5}$$

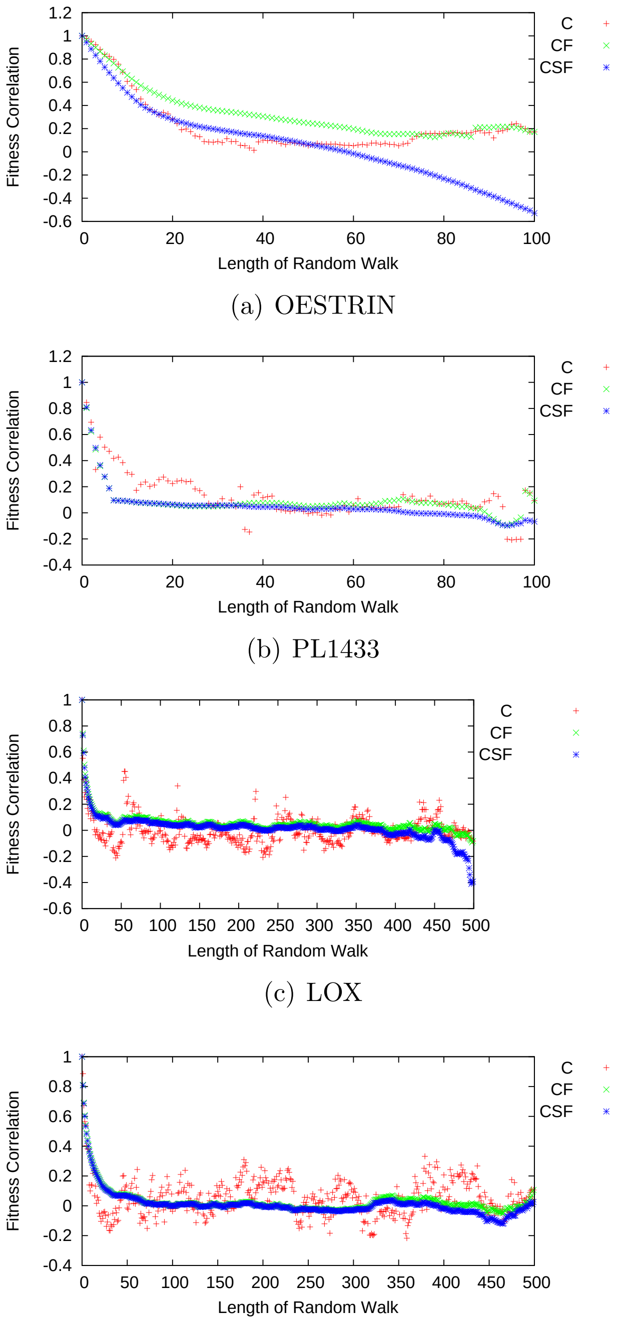
Moreover, we have (5.6) extended from the above (5.5):

$$\rho_s = \frac{\frac{1}{n+1}\sum_{i=0}^{m}\sum_{j=0}^{n}(x_{ij}-\mu_0')(x_{i(j+s)}-\mu_s')}{\sigma_0'\sigma_s'}, \tag{5.6}$$

The method (5.6) is not following a rigorous standard as we are actually having different populations of samples for each different lengths of random walk, i.e, the more steps in random walks, the more configurations to investigate.

Other than (5.6), we also adopt mean values and standard deviations from all configurations in the search space, as in (5.7):

$$\rho_s = \frac{\frac{1}{n+1}\sum_{i=0}^{m}\sum_{j=0}^{n}(x_{ij}-\mu)(x_{i(j+s)}-\mu)}{\sigma^2}, \tag{5.7}$$



(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.4: Fitness Correlations in Molecular Landscapes

The averaged autocorrelation values of (5.5), (5.6) and (5.7) are depicted in Figure 5.4, by C, CF, and CSF, respectively.

Clearly, in 5.4 the correlations are getting weaker along the path of random walks. Compared to the other three landscapes, the correlation curve drops fastest to 0 in Landscape PL1433, and subsequently remains prevailing flat, i.e. the correlation length in Landscape PL1433 is the smallest among these four molecular landscapes. In other words, Landscape PL1433 is the most rugged among the four. It is also easy to see that the more configurations involved in the analysis, the less noisy will it be. The curve of CSF almost coincide with the one CF. This provides evidence for the assumption on the statistically isotropic property of molecular landscapes, for random walks as being stationary process in the landscapes.

An extraordinary observation is the negative correlation CSF shows up in Landscape OESTRIN after a certain length of random walk. This means that the random walk starts from a peak but ends at the bottom, which should not happen in a strict random walk.

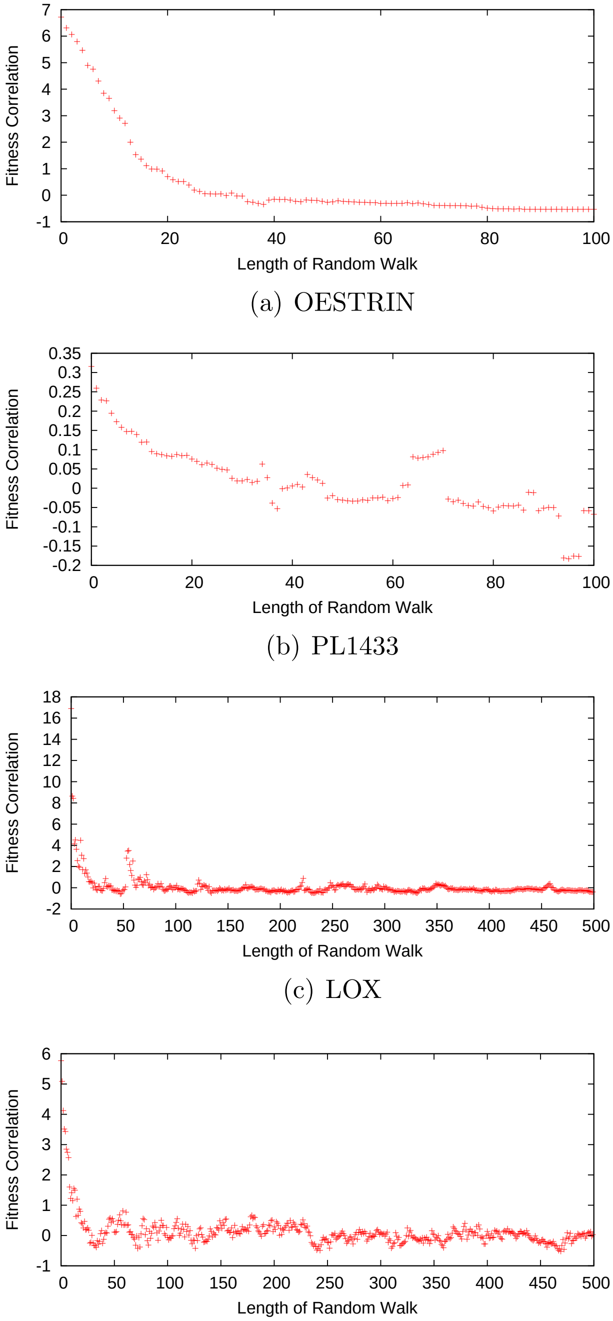
As we mentioned earlier in this section, applying the mean and standard deviation of all configurations to relatively small local populations may lead to high errors in estimation. (5.8) can function such a problem. It uses the mean and standard deviation of all configurations we obtained, but apply them only to the starting points (configurations in column $0, x_{i0}$) and the points x_{is} gained by s step random walks from x_{i0}.

$$\rho_s = \frac{\frac{1}{n+1}\sum_{i=0}^{n}(x_{i0}-\mu)(x_{is}-\mu)}{\sigma^2}. \tag{5.8}$$

Figure 5.5 illustrates the outcome of (5.8). The normal self-correlation coefficient ρ_0 (the correlation at step 0) should be 1 or approximately 1. But we can see in Figure 5.5 that, ρ_0 is approximately 7, 0.33, 9 and 6 in Landscape OESTRIN, PL1443, LOX, and NPY, respectively.

5.2 Information Analysis

As explained in section 3.2 and 3.3, both information analysis methods, information content and partial information content, apart from neutrality measures, are ϵ indicators. All of them are based on discretization processes



(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.5: Fitness Correlation

which transfer continuous models into discrete counterparts with state space $\{1, \bar{1}, 0\}$.

The scale of the values of the threshold ϵ in information content and partial information content is chosen based on its performance in the analysis of neutrality measures. This will be further discussed in section 5.3.

5.2.1 Information Content

For OESTRIN and Landscape PL1443, we have information content of five different random walks lengths $L = 20, 45, 60, 75, 100$, while for Landscape LOX and Landscape NPY, $L = 20, 50, 100, 250, 400, 499$. Information content $H(\epsilon)$ of these four landscapes is depicted in Figure 5.6.

As we can see in the Figure 5.6 that, $H(\epsilon)$ is an increasing function for low values of ϵ, where $H(\epsilon_1) < H(\epsilon_2)$ when $\epsilon_1 < \epsilon_2$. This observation confirms the conclusion that have been previously drawn in Vassilev et al. (2000). These landscape paths are characterized by relatively small flat landscape areas, since each path as an ensemble consists mainly of two types of objects. It is also declared that the flat landscape areas prevail over the ruggedness in a time series, if $H(\epsilon)$ is a decreasing function.

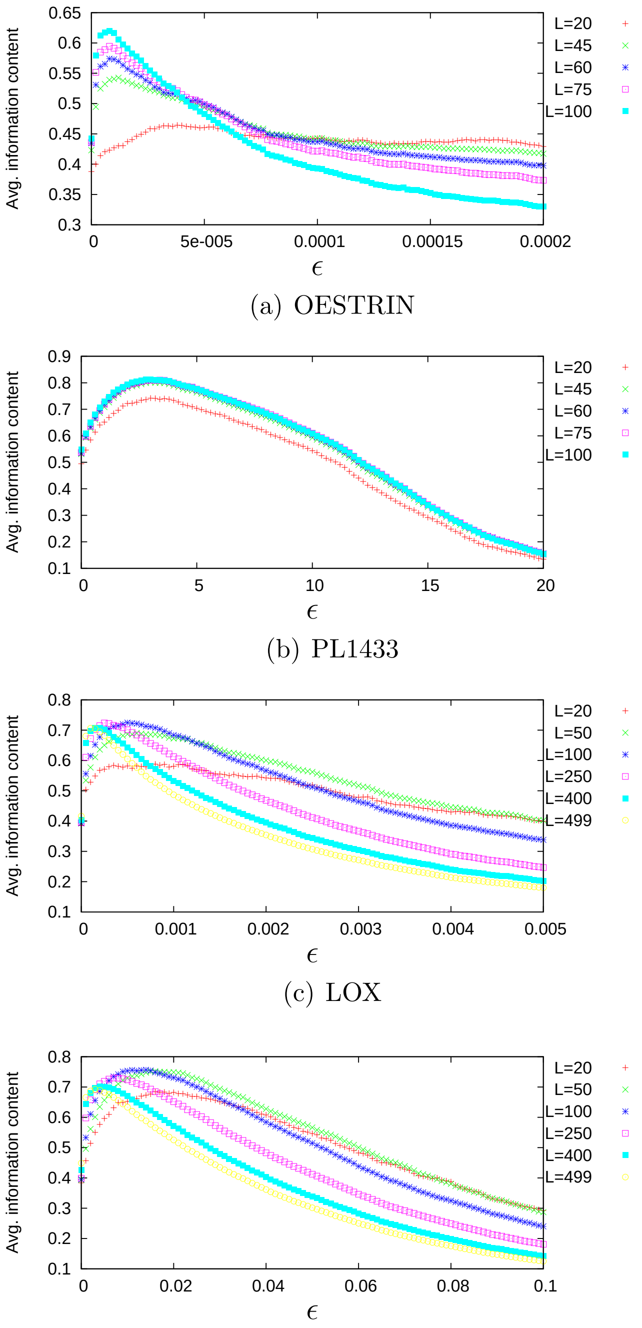
The graphs $(a), (c), (d)$ in 5.6 show that for large values of ϵ, the smaller the length of random walks, the slower information content $H(\epsilon)$ is descending. We thus can say that these landscapes are relatively smooth, and better correlated than Landscape PL1443, which is depicted in (b).

In the way of (5.6) and (5.7) for correlation analysis, we regard a configuration obtained along the random walk path as a starting point as well to its succeeding configurations. We apply this method to information content analysis as well, as in Figure 5.7.

Compared to 5.6, we can see in 5.7 that the areas where $H(\epsilon)$ is a decreasing function are significantly smaller. In other words, the search spaces are less rugged than in 5.6.

5.2.2 Partial Information Content

As in the case of information content analysis, for OESTRIN and Landscape PL1443, we have information content of five different random walks lengths



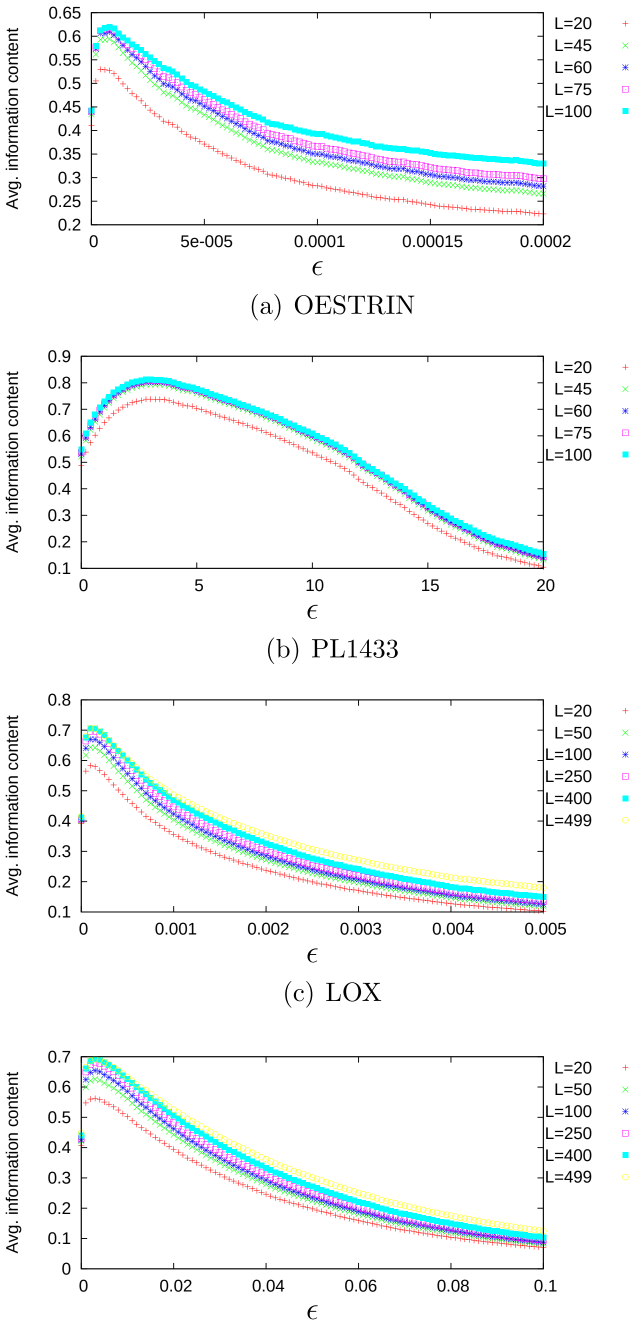
(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.6: Information Content



(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.7: Information Content

$L = 20, 45, 60, 75, 100$, while for Landscape LOX and Landscape NPY, $L = 20, 50, 100, 250, 400, 499$. Partial Information content $M(\epsilon)$ is depicted in Figure 5.8.

In 5.8, the partial information content decreases toward 0 as ϵ increases. Back to 3.2.2, the function of partial information content (3.1) defined by (Vassilev et al. 2000) has been mentioned as:

$$M(\epsilon) = \frac{\nu(\epsilon)}{n}$$

where ν is number of the optima along the random walk path. Id est, the steepness of $M(\epsilon)$ indicates the diversity of the optima when they are classified by their magnitude. Therefore we draw the conclusion from 5.8 that, the number of optima in Landscape PL1443 tops the other three landscapes. In other words, it will be more difficult to explore and estimate the entire Landscape PL1443 than the others.

As well as for information analysis, we also make use of the configurations by the means in (5.6) and (5.7), as it is plotted in Figure 5.9.

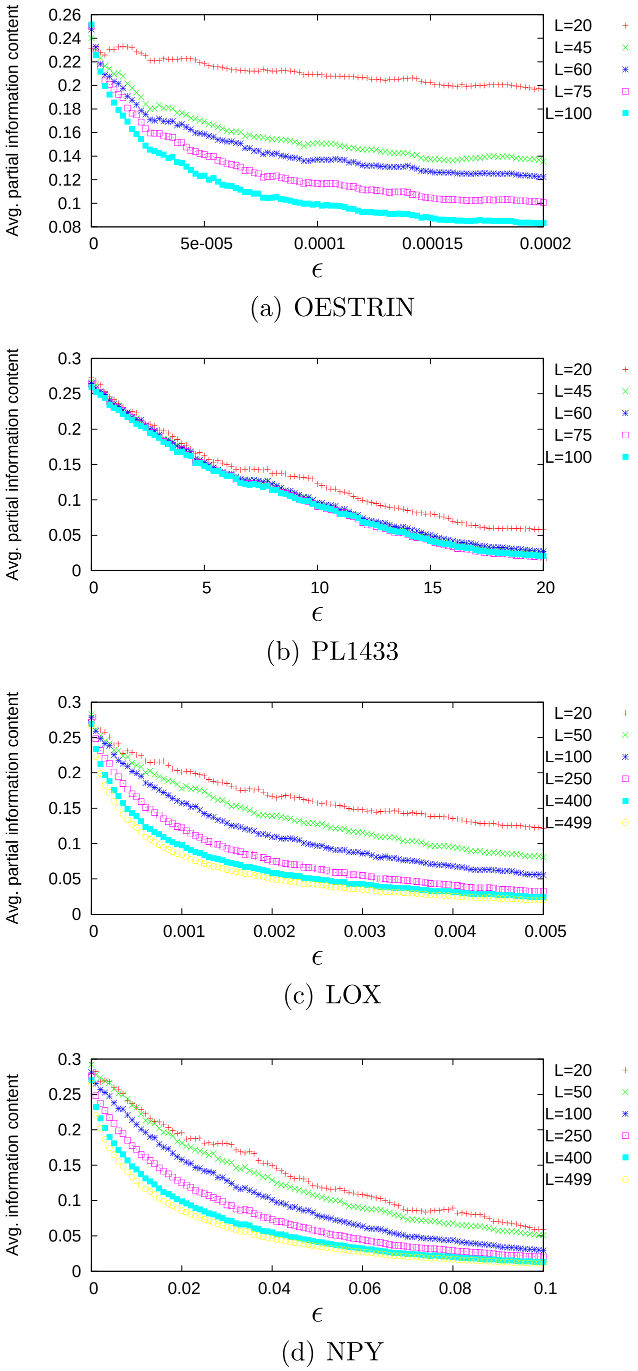
From 5.9 we can see that, with the random walk length $L >= 60$, the curves of $M(\epsilon)$ almost coincide with each other. We thus assume that $L = 60$ is a sufficient length to illustrate the modality of these four landscapes.

By comparing 5.9 to 5.8, we can also find that for the same length of random walks, the more configurations, the faster the curve of $M(\epsilon)$ declines. We cannot make a definite conclusion from this observation. We can only say that in these four landscapes, the increasing ratio of the number of local optima is smaller than the one of the number of configurations involved.

5.3 Neutrality Measures

The changes in the neutrality of these four fitness landscapes for increasing lengths of neutrality blocks and values of threshold ϵ are displayed in Figure 5.10. For both OESTRIN and Landscape PL1443, 100 steps of random walks are performed, while for Landscape LOX and Landscape NPY, 499 steps of random walks are performed.

As in Figure 5.10, the frequency of $S(\epsilon)$, which is the average lengths of neutrality blocks in our analysis, is noticeably approaching linear diagonal



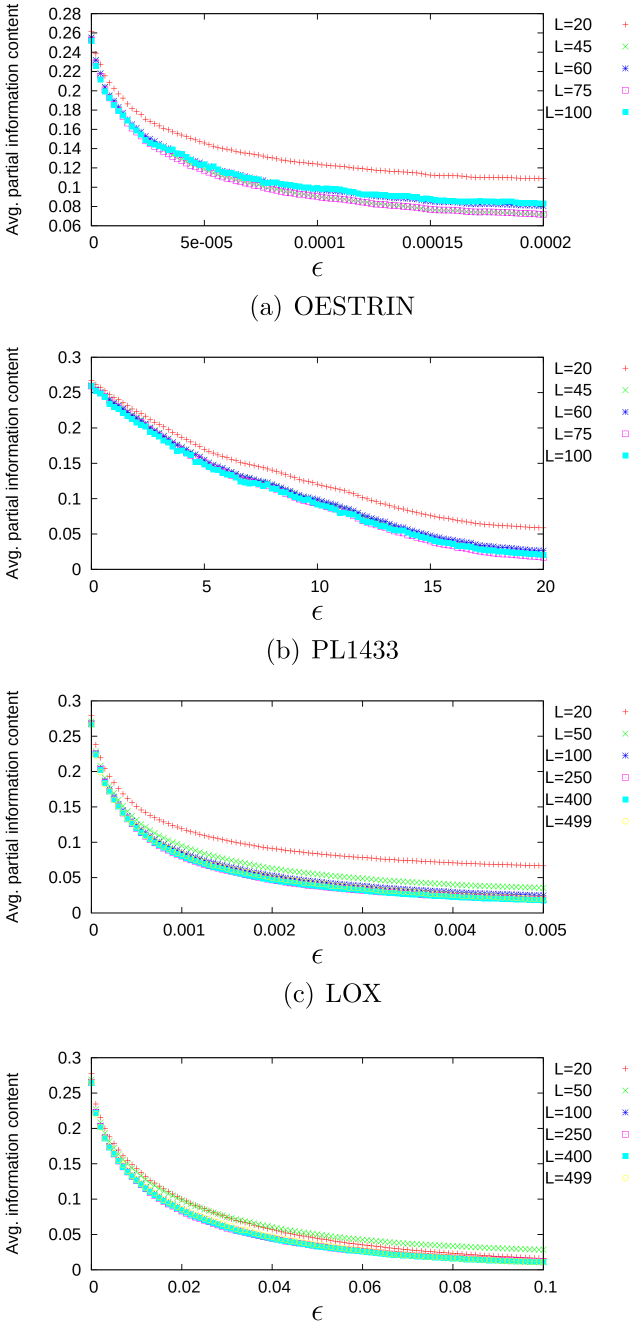
(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.8: Partial Information Content



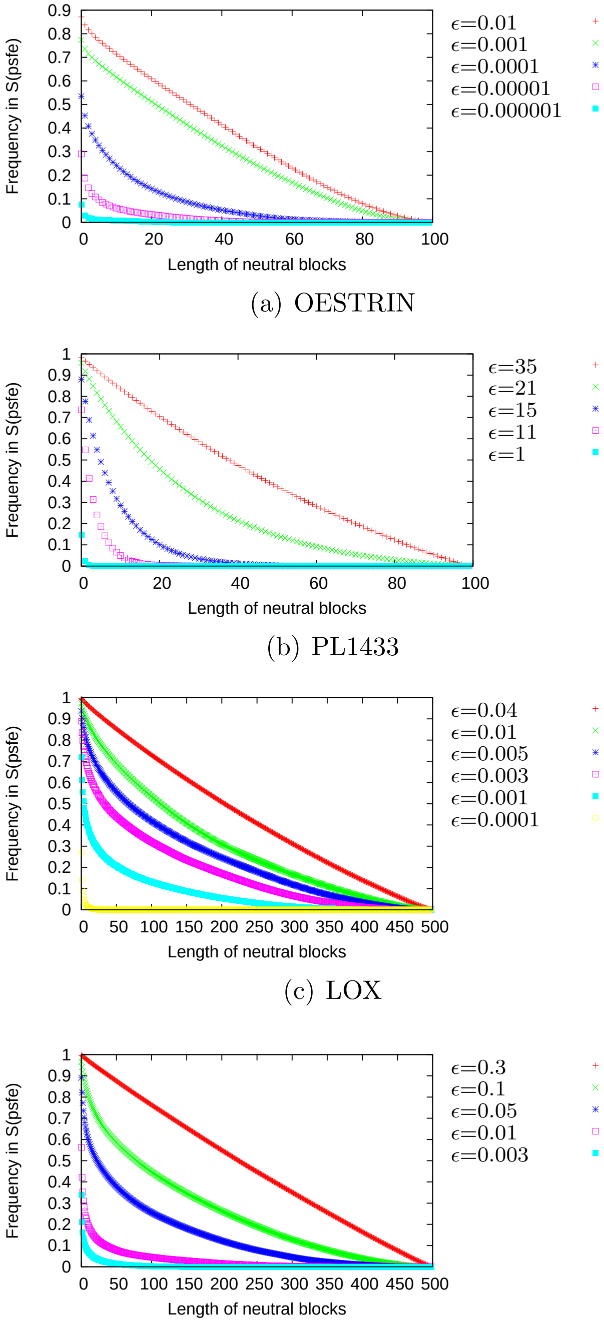
(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.9: Partial Information Content



(a) OESTRIN

(b) PL1433

(c) LOX

(d) NPY

Figure 5.10: Neutrality Measures

by increasing the value of ϵ, so will more interesting spots be missing. Per contra, when ϵ is too small, which makes $S(\epsilon)$ getting too close to 0, the exploration will be stuck by too many local optima.

The values of threshold ϵ which perform better in the analysis of neutral measures, will lead to a better performance as well in the analysis of both information content and partial information content. Since it is easier and more intuitive to obverse better solutions of ϵ in neutrality measures, the scale of ϵ applied to information analysis is set in consequence of its performance in neutrality measures.

Chapter 6

Conclusions and Outlook

Both the large population of molecules and the uncertainty of the transformations of a molecule contribute to the complexity and difficulty in molecular landscape analysis. Algorithms that can adapt to such situations are hence of crucial importance in the approach to good solutions. As a precursory step to applying these algorithms, our research examines molecular landscape characteristics to determine the difficulties search-algorithms will encounter when applied to the landscape.

After introducing the concept of a landscape, we described four measures for analysing landscapes: correlation analysis, information content, partial information content, and neutrality measures. We then applied these measures to four molecular landscapes: OESTRIN, PL1433, LOX, NPY, which represent the Landscape of Oestrogen Receptor, 14-3-3 γ Isoform Receptor, Lipoxygenases Inhibitor, and Neuropeptide Y_2 Receptor, respectively.

Correlation analysis reflects the linear interdependence of two molecular activities along a random walk. The lower the correlations, the more rugged the landscape. Landscape PL1433 hence turns out to be the most rugged among the four landscapes we investigated, whose correlation decreases to 0 most swiftly. The plots CF and CSF of (5.6) and (5.7) respectively make it clear that the more configurations are involved in the investigation, the less noisy the correlation will be. It also provides evidence for the assumption on the statistically isotropic property of molecular landscapes as CF and CSF almost coincide to each other in spite of the different numbers of analysed configurations. Nevertheless, if a relatively small subset is taken from a large set of configurations, then the mean and standard deviation of this subset may be remarkably different from the global mean and deviation. Ergo applying the global mean and deviation to this subset will result in high errors in estimation.

Another observation revealed in correlation analysis is that in Landscape OESTRIN, the random walk is actually not strictly defined since the starting points are not really randomly chosen but the ones with good fitness values.

The information content, partial information content, and neutrality measures are three threshold-based indicators which depict the ruggedness, modality and neutrality of a landscape respectively. The results of information content and partial information content confirms that the four molecular landscapes are relatively correlated and smooth. The comparisons among the four landscapes also correspond with the one we concluded from correlation analysis that Landscape PL1443 is the most rugged.

The neutrality measure in our research does not only represent the neutrality of the landscapes, but is also an intuitive means that yields a good threshold. A good threshold should not exhibit any extreme neutrality behaviour, i.e., it should not result in either the absence or abundance of neutral blocks.

Our research confirms that despite the large population of configurations, a molecular landscape properties can be estimated without full exploration of the search space by information characteristics analysis, which can be used to steer the further research, such as algorithm selections.

Recent research based on random walks conducted in molecular landscapes is normally not strictly random, for the starting points are usually chosen from those with good fitness values. This is an intuitive way to approach a good solution, however, it may also result in the risk of missing the interesting points which are closer to comparatively worse solutions. For future work, we thus propose to start the random walks from the boundaries between active "spots" in chemical space instead. There are several possible definitions of boundaries,

1. a cluster around a given active compound "C" for a given threshold "t" is a set of compounds with similarity to "C" greater than or equal to "t". A boundary is the in the intersection of two such clusters. We need to make sure that "t" is small enough so that there will actually be an intersection.

2. a starting point is one which has the same similarity(distance) to 2 active compounds, and (only if) is also further from the rest of active compounds. In this case, we do not need clusters, but only focus on similarities.

   Suppose we have 20 active compounds. Lets call the active compounds C_i with i between 1 and 20. The border between components C_1 and C_2 consists of those components x that are equally similar to C_1 and C_2, i.e., $s(x, C_1) = s(x, C_2)$, and more similar to C_1 and C_2 than to any other active compound, i.e., $s(x, C_1) = s(x, C_2) >= s(x, C_j)$ for j between 3 and 20.

3. the intersection of the above 2 definitions.

This idea helps to eliminate the bias in molecular landscapes exploration, and to enhance the chance of reaching good “hidden” solutions in the search space.

Acknowledgements

The author would like to thank her academic advisor Dr. Michael Emmerich for his support, guidance and patience. His comments and suggestions make this thesis much more readable. She would also like to thank her second supervisors, Dr. André Deutz and Dr. Andreas Bender for their technical advice and feedback on this work. In addition, the author would like to thank the following members from Pharma-IT group, Johannes Kruisselbrink, Eric Faddeev(H.S. Faddiev), and Alexander Aleman, for the data and assistance they offered during the research.

The author would like to thank her spiritual mentor Fr. Ben Engelbertink. In his wisdom and kindness, she can always find solace and strength, especially during those difficult moments. She is also grateful to her friends. They have made her time in Leiden delightful and worthwhile. Finally, the author would like to express her gratitude to her family, especially to her mother, Joe Cheung, and AnaSkula V., whose immense intellectual and emotional support she has been privileged bestowed.

Bibliography

Brooijmans, N. and I. Kuntz (2003). Molecular recognition and docking algorithms. *ANNUAL REVIEW OF BIOPHYSICS AND BIOMOLECULAR STRUCTURE 32*(146), 335–373.

Brown, N. (2009). Chemoinformatics—an introduction for computer scientists. *ACM Comput. Surv. 41*(2), 1–38.

Enmark, E. and J.-Å. Gustafsson (1999). Oestrogen receptors - an overview. *Journal of Internal Medicine 246*(2), 133–138.

Faddiev, H. S. (2008, December). Computer aided design of a ligand specific to 14-3-3 gamma isoform, danio rerio.

Garnier, J. and L. Kallel (2002). Efficiency of local search with multiple local optima. *SIAM J. Discret. Math. 15*(1), 122–141.

Jones, T. and S. Forrest (1995). Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In *Proceedings of the 6th International Conference on Genetic Algorithms*, San Francisco, CA, USA, pp. 184–192. Morgan Kaufmann Publishers Inc.

Kitchen, D. B., H. Decornez, J. R. Furr, and J. Bajorath (2004). Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Discov 3*, 935–949.

Korach, K., J. Couse, S. Curtis, T. Washburn, J. Lindzey, K. Kimbro, E. Eddy, S. Migliaccio, S. Snedeker, D. Lubahn, D. Schomberg, and E. Smith (1996). Estrogen receptor gene disruption: molecular characterization and experimental and clinical phenotypes. *Recent Prog Horm Res 51*, 159–86; discussion 186–8.

Kruisselbrink, J. W., A. Aleman, M. T. Emmerich, A. P. IJzerman, A. Bender, T. Baeck, and E. van der Horst (2009). Enhancing search space diversity in multi-objective evolutionary drug molecule design using niching. In *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, New York, NY, USA, pp. 217–224. ACM.

Kruisselbrink, J. W., T. Bäck, A. P. IJzerman, and E. van der Horst (2008). Evolutionary algorithms for automated drug design towards target molecule properties. In *GECCO '08: Proceedings of the 10th annual conference on Genetic and evolutionary computation*, New York, NY, USA, pp. 1555–1562. ACM.

Lameijer, E.-W., A. IJzerman, and J. Kok (2005). The molecule evoluator: an interactive evolutionary algorithm for designing drug molecules. In *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*, New York, NY, USA, pp. 1969–1976. ACM.

Leier, A. and W. Banzhaf (2003, 8-12 December). Exploring the search space of quantum programs. In R. Sarker, R. Reynolds, H. Abbass, K. C. Tan, B. McKay, D. Essam, and T. Gedeon (Eds.), *Proceedings of the 2003 Congress on Evolutionary Computation CEC2003*, Volume 1, Canberra, pp. 170–177. IEEE Press.

Michel, M. C., A. Beck-Sickinger, H. Cox, H. N. Doods, H. Herzog, D. Larhammar, R. Quirion, T. Schwartz, and T. Westfall (1998). XVI. International Union of Pharmacology Recommendations for the Nomenclature of Neuropeptide Y, Peptide YY, and Pancreatic Polypeptide Receptors. *Pharmacol Rev 50*(1), 143–150.

Palmer, P. (1991). Optimization on rugged landscapes. In A. S. Perelson and S. A. Kauffman (Eds.), *Molecular Evolution on Rugged Landscapes: Proteins, RNA, and the Immune System: the Proceedings of the Workshop on Applied Molecular Evolution and the Maturation of the Immune Response, Held March, 1989 in Santa Fe, New Mexico*, pp. 3–25. Addison Wesley Publishing Company.

Stadler, P. (2002). Fitness landscapes. In *Lecture Notes in Physics*, Volume 585/2002, pp. 183. Springer Berlin / Heidelberg.

Törn, A. and A. Žilinskas (1989). *Global Optimization*, Volume 350/1989. Springer Berlin / Heidelberg.

Vassilev, V. K., T. C. Fogarty, and J. F. Miller (2000, Spring). Information characteristics and the structure of landscapes. *Evolutionary Computation 8*(1), 31–60.

Weinberger, E. D. (1990). Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics 63*, 325–336.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*.