NAIST-IS-DD0461205

Doctoral Dissertation

Studies on Prediction of Protein Function Based on Oligopeptides

Hisayuki Horai

February 2, 2006

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to the Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Science.

Thesis Committee:

Professor Toshio Hakoshima	(Supervisor)
Professor Shigehiko Kanaya	(Co-supervisor)
Professor Hirofumi Doi	(Member)
Associate Professor Kouichi Doi	(Member)

オリゴペプチドに基づく
タンパク質機能予測に関する研究*

蓬莱尚幸

内容梗概

本論文では、オリゴペプチドに基づく新しいタンパク質予測手法について論じる。本論文の主たる目的は、オリゴペプチドの概念がさまざまなタンパク質機能に関する効率的な予測手法を開発するために有効であることを示すことである。本論文で提案する予測手法では、あるタンパク質の既知の機能は、そのオリゴペプチドに継承され、オリゴペプチドと機能の関連度を全タンパク質に対して計算され、この関連度を用いて任意のタンパク質の機能を自動的に予測する。

本論文では、ヒトの標準的なタンパク質を利用して、さまざまな酵素活性および GeneOntology用語に対する提案手法による予測結果を計測し評価する。本論文での予測精度評価には、情報検索分野で提案され通常的に利用されている手法を用いる。これらの予測実験により、提案する手法は多くのタンパク質機能に関して非常に有効であることが示唆される。

提案する手法の予測性能を客観的に示すために、本論文では、既存の予測手法である相同性検索およびパターンマッチングと比較する。対照的予測実験により、提案する手法はパターンマッチングより非常に有効であり、相同性検索に対しても同等以上の性能を持つことが示唆される。

オリゴペプチドの長さは、提案する手法における重要なパラメータであり、予測性能に影響を与える。本論文では、長さ１から９のオリゴペプチドについて、それらを用いた場合の予測性能を比較する。長さ1のオリゴペプチドには予測能力はなく、長さ５以上ではほぼ同等であることが明示された。また、長さ２から４の予測性能は、機能により異なるが、いずれの場合も長さと予測性能は正の相関を示すことも判明した。また、オリゴペプチドが長くなるにつれ、共有性は低下し、予測手法適用性の低下を招く。オリゴペプチド共有性の評価を考慮に加えると、任意のタンパク質機能の予測に対して汎用的に利用すべきオリゴペプチドの長さは５または６であることが示唆される。

さらに、オリゴペプチドと機能との関連度を利用して、機能に特有かつ一般的なオリゴペプチドを提示する手法を例示する。

キーワード

タンパク質、機能予測、オリゴペプチド、酵素活性、GeneOntology

*奈良先端科学技術大学院大学情報科学研究科情報システム学専攻学位論文, NAIST-IS-DD0461205, 2006年 2 月 2 日.

Studies on Prediction of Protein Function Based on Oligopeptides*

Hisayuki Horai

Abstract

This thesis proposes and investigates a new prediction method of protein function based on oligopeptides. The main purpose of the thesis is to demonstrate that `oligopeptide' enable us to develop an effective method for predicting various protein function. In the proposed method, a known function of each protein is regarded to be inherited to its oligopeptides, the correspondence between an oligopeptide and a function is calculated in the whole proteins, and unknown functions of an arbitrary protein are predicted by means of the correspondence automatically.

The prediction performance of the proposed method is measured and evaluated through several experimental predictions for functions including enzyme activities and GeneOntology terms using the whole human proteins. In order to evaluate the performance of prediction, the thesis utilises evaluation methods proposed and commonly used in the research domain of information retrieval. The results of the comparative studies suggest that the proposed method is quite efficient for various protein functions.

The thesis also characterises the relation between the length of oligopeptides and the prediction of protein functions. The performance of prediction is measured for the length of oligopeptides between 1 and 9. The results suggest that oligopeptides of the length of 1 has no predictability, oligopeptides longer than 4 are almost equally effective for all functions. The predictability of oligopeptides of the length between 2 and 4 depends upon the functions, and the longer oligopeptides are more efficient than the shorter ones for each function. Furthermore, the longer oligopeptides are more versatile than the shorter one because the longer oligopeptide is more varied than the shorter one and the degree of the coexistence is inversely related to the length. Considerations on statistics of oligopeptides suggest that the most acceptable length of oligopeptides is 5 or 6 of generally predicting an arbitrary function.

The thesis also describes an example of finding oligopeptides which are specific to and general in a function by means of the correspondence between each oligopeptide and a function

Keywords:

Protein function, prediction, oligopeptide, enzyme activity, and GeneOntology.

.*Doctoral Dissertation, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0461205, February 2, 2006.

Contents

1. Introduction

Prediction is researched widely in bio-informatics [9-27]. Proteomics is the large-scale study of proteins, particularly their structures and functions, and their existence. Identifying the function of each newly determined sequence by means of bioinformatics techniques is one of the most important problems in proteomics [5,6,57]. Although the function of each protein can be introduced in wide spectrum and predicted based on different properties, the thesis focuses on the prediction of protein functions based on sequence. There are some methods for solving the problem proposed based on homology search [6] and pattern matching [5].

Each method based on homology search focuses on similarity of a relatively long subsequence or the full-length sequence. In many cases, each protein is related to several numbers of functions. When a new protein has homology to such a multi functional protein, it is difficult to determine that each function is annotated or not. Consequently, further investigation of a protein at the level of every shorter subsequence is needed after homology search.

Each method based on pattern matching focuses on similarity of a relatively short subsequence. These are conservative methods, taking similarity to clearly defined protein families whose members are annotated with functions. It is difficult to predict all functions because many functions have not been able to relate with any families yet.

Oligopeptide is a subsequence of fixed length. For example, in the 28,520 whole human proteins registered in RefSeq (Reference Sequence of the National Center for Biotechnology Information dated 13-May-2005), there exist 2,361,750 kinds of oligopeptides of length 5 [3, 33]. The existence of oligopeptides shows quite interesting characteristics [2] : (1) some oligopeptides exist commonly in many proteins and others exist unevenly; (2) some oligopeptides exist too many time in comparison with the existing probability of each component amino acid; and (3) many oligopeptides do not exist in the world of proteins (specificity of oligopeptide). Therefore, to view the world of proteins from the perspective of oligopeptides will provide a new computational science of proteomics. As one of the first steps of such computational proteomics from the perspective of oligopeptides, the thesis proposes a new method based on the concept of the lexicon of oligopeptides that have been paid much effort to construct by some researchers [2]. In our method, each protein is characterised based on the existence of oligopeptides.

In a large part of this thesis, oligopeptide of length 5 is utilised. The length of 5 is not mandatory in the proposed method. The discussion on the length of oligopeptides also appears in this thesis. Our method is based on the co-occurrence of each oligopeptide and the specificity of oligopeptide. Longer the length of oligopeptide is, less the co-occurrence of each oligopeptide, while, in [5], the specificity of oligopeptide does not be observed strongly in oligopeptides whose length is less than 5. The thesis discuss the matter more precisely by some experiments and observation of the results using a real set of whole human proteins.

Our method predicts the functions annotated to a protein based on a set of proteins already annotated, called Annotated Proteins here. Every Annotated Protein is divided into a set of its oligopeptides, and each function annotated to the protein is regarded to be related to all of its oligopeptides. Finally, the correspondence between oligopeptides and functions in Annotated Proteins is calculated. The correspondence between an oligopeptide and a function is the number of proteins which contain the oligopeptide and be annotated with the function. This correspondence is uniquely defined for each set of Annotated Proteins and stored in a vector, PepFunc Vector.

The correspondence between a new protein and each function is calculated based on all oligopeptides in the protein and PepFunc Vector.

The thesis evaluates the prediction performance through several experiments. In the evaluation, the thesis utilises some measurements used in information retrieval research, such as recall precision and f-measure [4]. In biolinguistics, i.e. information retrieval research sub-domain of bioinformatics, the measurements are commonly used [28-32]. These measurements are effective to evaluate a score-based prediction method. Our method is regarded as a score-based method using the correspondence as score. In a score-based method, the global property of performance for varied score threshold is more important than the best performance by a specific threshold. The global property is usually shown in a recall precision graph.

2. Prediction Method

This section explains a method proposed and discussed in the thesis to predict a function of a protein from its sequence based on oligopeptides.

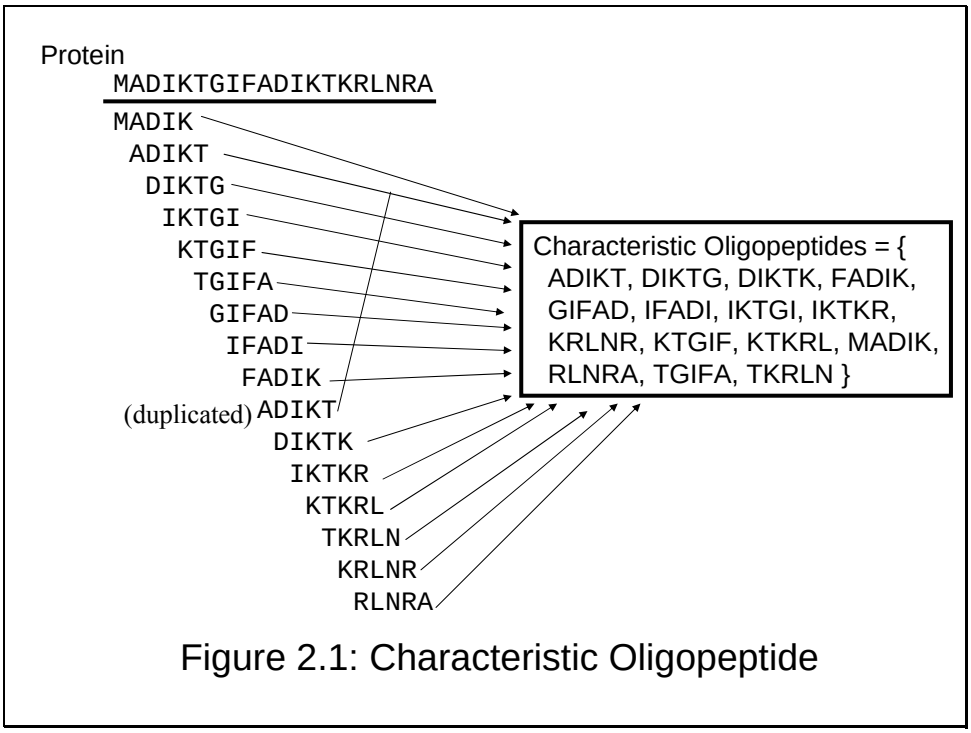

Figure 2.1: Characteristic Oligopeptide

2.1 Characteristic Oligopeptide

In the proposed method, each protein is characterised by a set of oligopeptides, called Characteristic Oligopeptides. The length of an oligopeptide is arbitrary fixed number n_{oligo}. Characteristic Oligopeptides of a protein are a set of all oligopeptides (without duplication) which exist in the protein. When the length of a protein is m then the number of Characteristic Oligopeptides is less than or equal to $m - n_{oligo} + 1$. If there is no duplication of oligopeptides in the protein, the number of its Characteristic Oligopeptides is equal to $m - n_{oligo} + 1$. Figure 2.1 shows a simplified small example for explanation. In this example, a set of Characteristic Oligopeptides is generated from a protein, whose sequence is MADIKTGIFADIKTKRLNRA and m is 20. In the example, n_{oligo} is 5, the only oligopeptide ADIKT appears twice, and 15 Characteristic

Oligopeptides are obtained for the protein.

2.2 Training Set for Prediction

For the prediction of an arbitrary function of protein *f*, the proposed prediction method needs a training set of proteins which is annotated appropriately concerning *f*. This means that some proteins are annotated that they have function *f* and others are not. This situation is quite common in any research domain where some automatic prediction methods are adopted to solve some problem.

The training set of proteins, the subset of proteins which are annotated that they have *f*, and the remaining subset of proteins which are not annotated are called as Annotated Proteins, Positive Proteins, and Negative Proteins, respectively.

2.3 PepFunc Vector

When an arbitrary function of protein *f* and Annotated Proteins *Ps* are given, the proposed method calculates the correspondence between *f* and all oligopeptides which appear in *Ps*. The correspondence is denoted by PepFunc Vector. PepFunc Vector is a real number vector. Each element is related to an oligopeptide and denotes the correspondence between *f* and its related oligopeptide, whose value is larger than or equal to 0.0, and smaller than or equal to 1.0. The order of the elements in PepFunc Vector is arbitrary because the prediction method does not focus on relation among oligopeptides but only on the correspondence between the function and oligopeptides. In other words, an element of PepFunc Vector is not suffixed by an integer which denotes the position in the vector but by the related oligopeptide itself. This subsection explains the calculation to generate PepFunc Vector from a set of Annotated Proteins *Ps*. PepFunc Vector generated from *Ps* and the element of PepFunc Vector related to oligopeptide *o* are represented by *VEC(Ps)* and *VEC(Ps)[o]*, respectively.

At first, the total union set of Characteristic Oligopeptides of all Annotated Proteins *Ps*, called Oligopeptide Universe of *Ps*, is generated. For each Annotated Protein *P*, i.e. an element of *Ps*, Characteristic Oligopeptides of *P*, represented by *OP(P)*, is obtained in the manner mentioned in the previous subsection. Oligopeptide Universe of *Ps*, represented by *OLIGO(Ps)*, is generated by the following equation:

$$OLIGO(Ps) = \bigcup_{P \in Ps} OP(P)$$

Each element of *OLIGO(Ps)* is related to an element of PepFunc Vector. The number of *OLIGO(Ps)* is the length of PepFunc Vector.

In the next, the number of Annotated Proteins and the number of Positive Proteins which obtain each oligopeptide of *OLIGO(Ps)* are counted. The number of Annotated Proteins and the number of Positive Proteins which obtain oligopeptides *o* are represented by $N_{all}(o)$ *and* $N_{positive}(o)$, respectively.

Finally, the correspondence between oligopeptide *o* and function *f*, i.e. *VEC(Ps)[o]* is calculated in the following equation:

$$VEC(Ps)[o] = \frac{N_{positive}}{N_{all}}, \text{ where } o \in OLIGO(Ps)$$

VEC(Ps)[o] is a real number between 0.0 and 1.0 because N_{all} is a positive integer, $N_{posotive}$ is a positive integer or 0, and N_{all} is greater than or equal to $N_{positive}$.

2.4 Prediction of Function

In the proposed method, the prediction of function *f* is the calculation of correspondence between an arbitrarily given protein *X* and a function *f* by means of preliminarily calculated PepFunc Vector *VEC(Ps)* by voting method. The correspondence is represented by *Cor(X, Ps)*.

A protein consists of many oligopeptides and each oligopeptide votes to judge whether the protein has the function or not. For every oligopeptide *o*, *VEC(Ps)[o]* is utilised as the point to vote. It means that the voting by *o* is weighted by the correspondence between *o* and *f*.

At first, for each Characteristic Oligopeptide *o* of the protein *X*, *VEC(Ps)[o]* is calculated. If *o* does not appear in any Annotated Protein, then *VEC(Ps)[o]* have not been calculated yet. In this case, *VEC(Ps)[o]* is defined as 0.0 because there is no Positive Proteins whose Characteristic Oligopeptides include *o*.

In the next, they are summed up for each occurrence of each Characteristic Oligopeptide *o* of the protein *X*. The summation is not for each Characteristic Oligopeptide but for each occurrence of each Characteristic Oligopeptide. If an oligopeptide appears in *X* more than once, *VEC(Ps)[o]* is multiplied by the number of occurrence and summed up.

Finally, the total summation is normalised by the number of oligopeptides in *X*, and *Cor(X, Ps)* is obtained. The number of oligopeptides in *X*, precisely speaking the number of all occurrence of all Characteristics of *X*, is $m - n_{oligo} + 1$, where m and n_{oligo} are the length of *X* and the length of an oligopeptide, respectively.

The calculation process results in the calculation of *Cor(X, Ps)* by the following equation, where *occur(X, o)* is the number of occurrence of oligopeptide *o* in *X*:

$$Cor(X, Ps) = \frac{1}{m - n_{oligo} + 1} \sum_{o \in OP(X)} \left(occur(X, o) \cdot VEC(Ps)[o] \right)$$

Cor(X, Ps) is a real number between 0.0 and 1.0 because *VEC(Ps)[o]* is a real number between 0.0 and 1.0 and *Cor(X, Ps)* is the arithmetic average of *VEC(Ps)[o]* for all occurrence of all Characteristic Oligopeptides in *X*.

For the practical use of the proposed method, i.e. the answering to YES/NO question whether a given protein has the function *f* or not, the threshold of the correspondence in order to divide any proteins into YES groups and NO groups according the value of correspondence is usually introduced. If the threshold is larger, then a smaller number of proteins belong to YES group. The decision of the threshold is recognised as one of the severe problems in the research domain of informal retrieval. The most appropriate value of the threshold is case by case: the threshold should be relatively low if a strong conservative prediction is purposed, while it should be relatively high if a week screening/filtering is purposed. In this thesis, the absolute value of the correspondence including an appropriate threshold is not out of scope, and the performance of the method is evaluated relatively and globally. This standpoint is quite common in the research domain of information retrieve. In information retrieve, some evaluation methods are proposed and utilised for score-base prediction system like the proposed method. The thesis evaluates the proposed method using such an evaluation method. This is discussed in the next section.

3. Evaluation Method and Materials

This section explains the evaluation method of the proposed method. The evaluation method is utilised throughout this thesis. The evaluation method needs some materials. This section also explains the materials.

3.1. Evaluation Method

In all experiments described in this thesis, the evaluation of the proposed method is performed by so-call 'Jack Knife Method' using a given set of proteins. The set of proteins are called as Protein Universe. For each protein in Protein Universe, the sequence must be fixed completely and it must be annotated whether it has function *f* or not.

For every protein *X* of Protein Universe, Protein Universe is divided into protein *X* and the remains of Protein Universe. The remaining proteins of Protein Universe except *X* are utilised as Annotated Proteins, PepFunc Vector is generated from the Annotated Proteins, and the correspondence between *X* and *f* is calculated. The prediction of *f* for *X* is performed by means of the relation between the other proteins and *f*. In other words, the obtained correspondence depends upon other proteins than *X* only, and *X* itself does not affect to the correspondence.

This prediction is performed for every protein in Protein Universe. The evaluation method consists of the repeated predictions by means of slightly different Annotated Proteins as many times as the number of proteins in Protein Universe.

After the predictions for all proteins in Protein Universe, the characteristics of prediction performance is evaluated by means of some measurements utilised in the research domain of information retrieval, such as precision (or 'accuracy'), recall (or 'sensitivity') and f-measure (i.e. harmonic average of precision and recall) as follows:

$$\text{Precision } P = \frac{TP}{TP + FP}$$

$$\text{Recall } R = \frac{TP}{TP + FN}$$

$$\text{F - measure } f = \frac{2 \times P \times R}{P + R}$$

TP = number of true positive

FP = number of false positive
FN = number of false negative

In order to overcome the threshold problem mentioned in the previous section, some evaluation methods are proposed and utilised in the research domain of information retrieve. The thesis evaluates the proposed method using such an evaluation method: calculate recall and precision for every threshold and drawing a recall precision graph.

To draw a recall precision graph, at first, Protein Universe is sorted in descending order, i.e. from the highest score to the lowest one. For each *i* from 1 to the number of Protein Universe, the top *i* proteins are selected. The top *i* proteins is regarded as a tentative prediction using *Cor(*the *i*-th protein*, Ps)* as threshold. Using each tentative prediction, the precision, the recall and the f-measure are calculated, and the values are plotted as a point in the recall precision graph. Finally, points as many as the number of Protein Universe are plotted in the recall precision graph, and these points are connected. The maximum f-measure is also selected.

3.2 Protein Universe

All evaluations of the proposed method throughout this thesis are performed by means of Protein Universe made from proteins in NCBI Reference Sequence (RefSeq) which provides a non-redundant set of proteins. The utilised version of RefSeq is dated 13-May-2005. Protein Universe includes all human proteins in RefSeq whose sequences are completely known, i.e. each of them does not include 'B', 'J', 'O', 'U', 'X' nor 'Z'. The number of Protein Universe is 28,520. 2,361,750 kinds of oligopeptides whose length is 5 are extracted from Protein Universe.

Proteins in RefSeq are annotated in GenBank format. All experiments are carried out on the assumption that RefSeq is annotated correctly and exhaustively. Based on the assumption, all subsets of Protein Universe made from RefSeq are utilised as Annotated Proteins which is divided into Positive Proteins and Negative Proteins.

4. Prediction of Enzyme Activities

In this section, the performance of the proposed method is evaluated for several enzyme activities. An enzyme activity is known as one of the important function of protein. The length of oligopeptide is a parameter of the proposed method and has an impact to the performance of the prediction. In this section, oligopeptide of length 5 is utilised.

4.1 Enzyme Activity and its Annotation to Protein

Enzymes activities are hierarchically classified and maintained by means of EC numbers [7, 34]. The EC number of an enzyme activity in the lowest class consists of 4 integers. For instance, EC2.7.1.112 denotes protein-tyrosine kinase [35, 36, 37]. A higher class of enzyme activity is denoted in the manner that the corresponding integer is replaced to character '-'. Character '-' mans 'any'. For instance, the parent class of EC2.7.1.112 is denoted as EC2.7.1.- and EC2.-.-.- denote the great-grand parent of EC2.7.1.112. There are six largest classes of enzyme activities from EC 1.-.-.- to EC 6.-.-.-.

Table 4.1: Number of Positive Proteins

EC number	Enzyme activity	Number of annotated proteins
EC2.7.1.112	Protein-tyrosine kinase	134 (0.5 %)
EC2.7.1.-	Phosphotransferases with an alcohol group as acceptor	424 (1.5 %)
EC2.7.-.-	Transferring phosphorous-containing groups	516 (1.8 %)
EC2.-.-.-	Transferases	862 (3.0 %)
EC1.-.-.-	Oxidoreductases	312 (1.1 %)
EC3.-.-.-	Hydrolases	813 (2.9 %)
EC4.-.-.-	Lyases	89 (0.3 %)
EC5.-.-.-	Isomerases	76 (0.3 %)
EC6.-.-.-	Ligases	146 (0.5 %)
EC-.-.-.-	(Enzyme)	2,260 (7.9 %)

Proteins in RefSeq are annotated with EC numbers. In the evaluation for a higher class of enzyme activity, Positive Proteins consists of proteins annotated as not

only itself but also its descendant. For instance, Positive Proteins for EC2.7.1.- are all proteins annotated as EC2.7.1.A and EC2.-.-.- are all proteins annotated as EC2.X.Y.Z where A, X, Y and Z are arbitrary integers or character '-'. Table 4.1 shows the number of Positive Proteins for all enzyme activities evaluated in this thesis. A percentage in Table 4.1 denotes the ratio to the number of Protein Universe.

EC2.7.1.112 is the enzyme activity which has the largest number of Positive Proteins among the lowest enzyme activities. Hierarchical enzyme activities from EC2.7.1.112 to EC2.-.-.- are evaluated in order to investigate the relation between the level of enzyme activities and the performance of the proposed method. Six largest classes of enzyme activities are evaluated in order to investigate the relation between the type of enzyme activities and the performance of the proposed method. EC2.-.-.- does not exist actually but generated in the thesis for the evaluation of the predictability for whole enzyme activities.

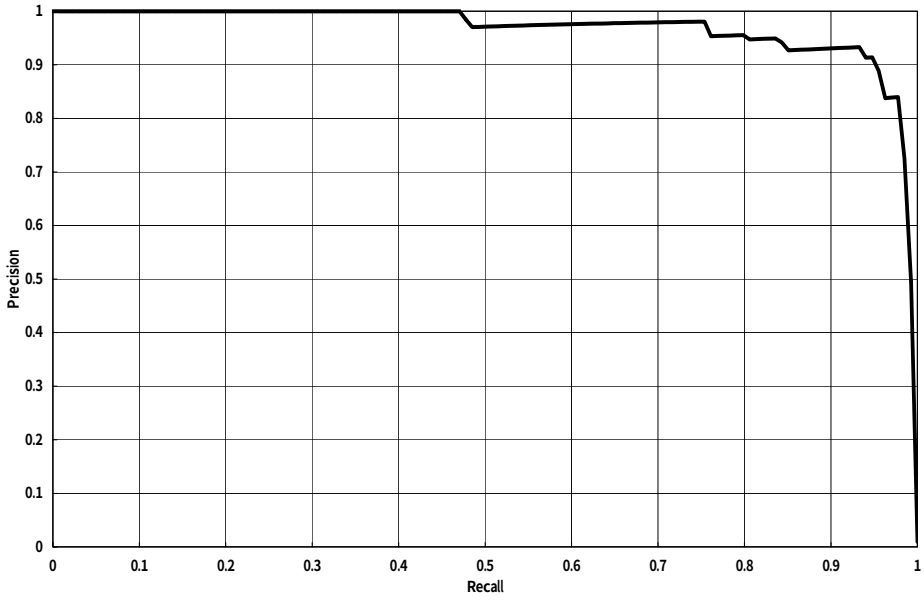

Figure 4.1: Recall Precision Graph for Protein-Tyrosine Kinase

4.2 Prediction of Protein-Tyrosine Kinase

This subsection describes the evaluation of the prediction of protein-tyrosine kinase (EC2.7.1.112) [35, 36, 27]. This is the enzyme activity which has the largest number of Positive Proteins among the lowest enzyme activities.

The recall precision graph of the prediction is shown in Figure 4.1. 63 (47.0%) proteins annotated with the EC number have been correctly predicted. The 50%, 80%, and 90% of proteins annotated with the EC number have been predicted with 97.1%, 94.7% and 93.0% precision, respectively. With 80% precision, the 98.5% of proteins annotated with the EC number have been predicted. The maximum f-measure is 0.932.

4.3 Prediction of EC2.7.1.-

This subsection describes the evaluation of the prediction of the parent class of protein-tyrosine kinase, EC2.7.1.- phosphotransferases with an alcohol group as acceptor. The enzyme activity includes protein kinase, diacylglycerol kinase, pantothenate kinase, hexokinase, galactokinase, and so on [38].

The recall precision graph of the prediction is shown in Figure 4.2. 116 (27.4%) proteins annotated with the enzyme activity and its descendants have been correctly predicted. The 50%, 80%, and 90% of proteins annotated with the enzyme activity and its descendant have been predicted with 96.4%, 85.8% and 65.8% precision, respectively. With 80% precision, the 87.3% of proteins annotated with the enzyme activity and its descendant have been predicted. The maximum f-measure is 0.839.

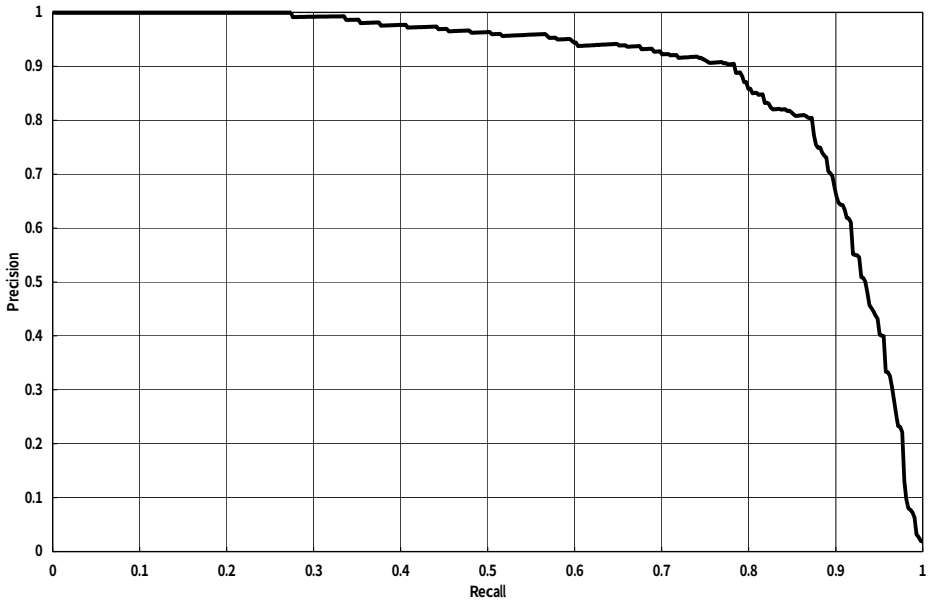

Figure 4.2: Recall Precision Graph for EC2.7.1.-

4.4 Prediction of EC2.7.-.-

This subsection describes the evaluation of the prediction of the grand parent class of protein-tyrosine kinase, EC2.7.-.- Transferring phosphorous-containing groups. The enzyme activity includes phosphotransferases, diphosphotransferases, nucleotidyltransferases and transferases for other substituted phosphate groups [39].

The recall precision graph of the prediction is shown in Figure 4.3. 122 (23.6%) proteins annotated with the enzyme activity and its descendants have been correctly predicted. The 50%, 66%, and 80% of proteins annotated with the enzyme activity and its descendant have been predicted with 95.9%, 93.2% and 82.1% precision, respectively. With 80% precision, the 82.9% of proteins annotated with the enzyme activity and its descendant have been predicted. The maximum f-measure is 0.822.

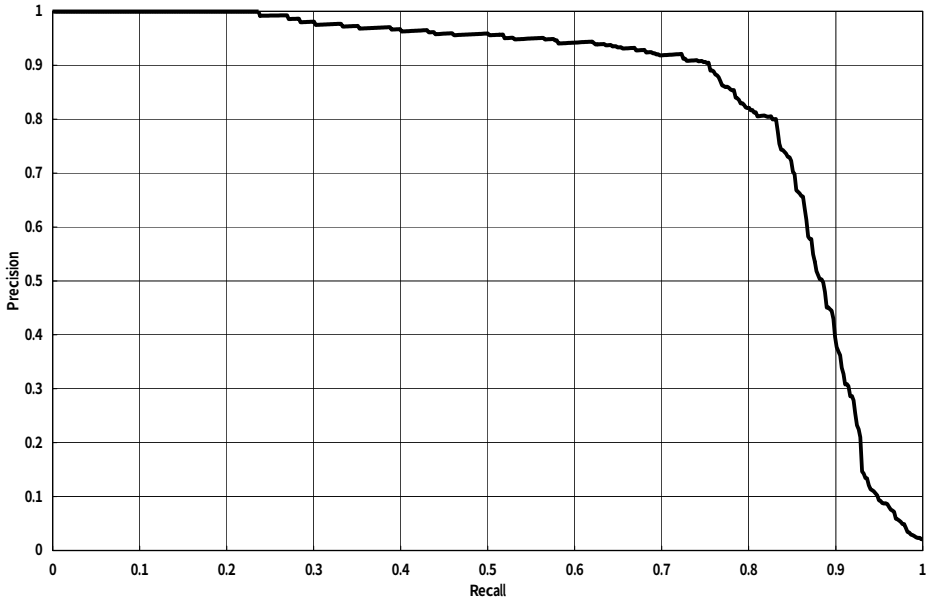

Figure 4.3: Recall Precision Graph for EC2.7.-.-

4.5 Prediction of Transferases

This subsection describes the evaluation of the prediction of transferases (EC 2.-.-.-). The enzyme activity is versatile including phosphotransferase, methyltransferase, acyltransferase, glycosyltransferase, transaminase, sulfurtransferase, and so on [40, 41].

Many kinases including protein-tyrosine kinase mentioned above are also classified in transferases.

The recall precision graph of the prediction for transferases is shown in Figure 4.4. 171 (19.8%) proteins annotated with transferases have been correctly predicted. The 50%, 66%, and 80% of proteins annotated with transferases have been predicted with 95.1%, 90.7% and 64.8% precision, respectively. With 80% and 50% precision, the 76.6% and 82.9% of proteins annotated with transferases have been predicted, respectively. The maximum f-measure is 0.786.

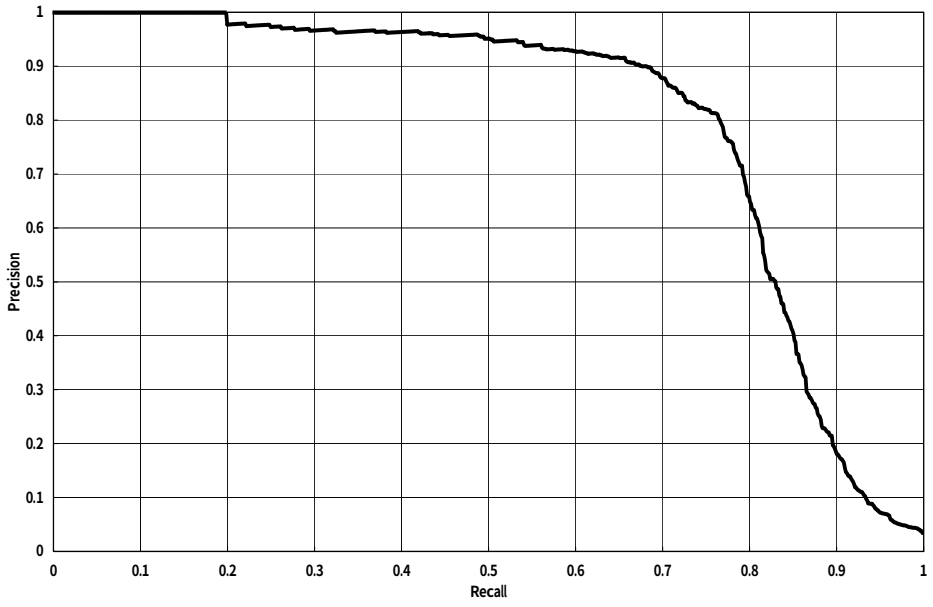

Figure 4.4: Recall Precision Graph for Transferases

4.6 Prediction of Oxidoreductases

This subsection describes the evaluation of the prediction of oxidoreductases (EC 1.-.-.-). The enzyme activity includes various oxidoreductases with many kinds of acceptors and donors [42, 43].

The recall precision graph of the prediction for oxidoreductases is shown in Figure 4.5. 39 (12.5%) proteins annotated with oxidoreductases have been correctly predicted. The 40%, 50%, and 66% of proteins annotated with oxidoreductases have been predicted with 89.2%, 76.8% and 17.2% precision, respectively. With 80% precision, the 45.6% of proteins annotated with oxidoreductases have been predicted.

The maximum f-measure is 0.630.

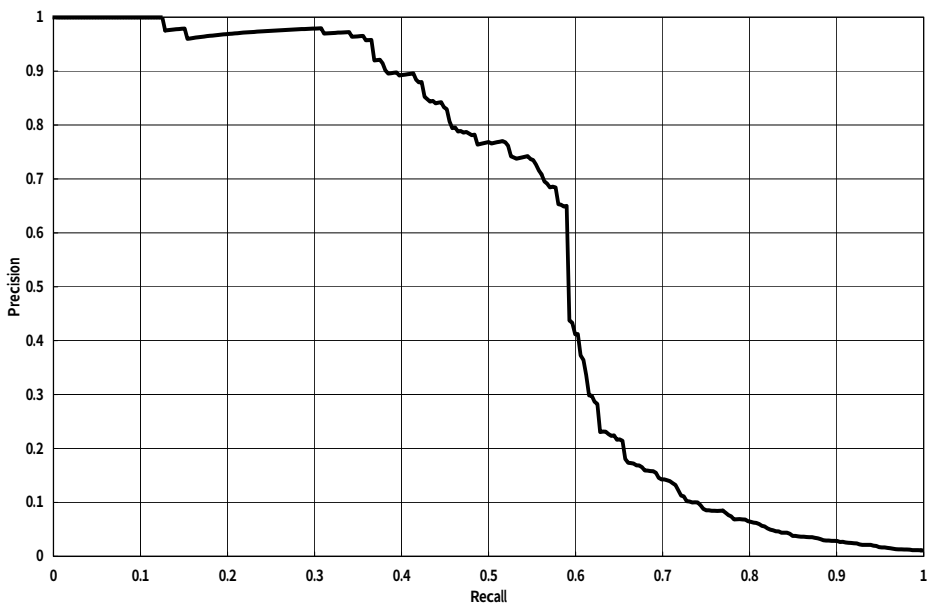

Figure 4.5: Recall Precision Graph for Oxidoreductases

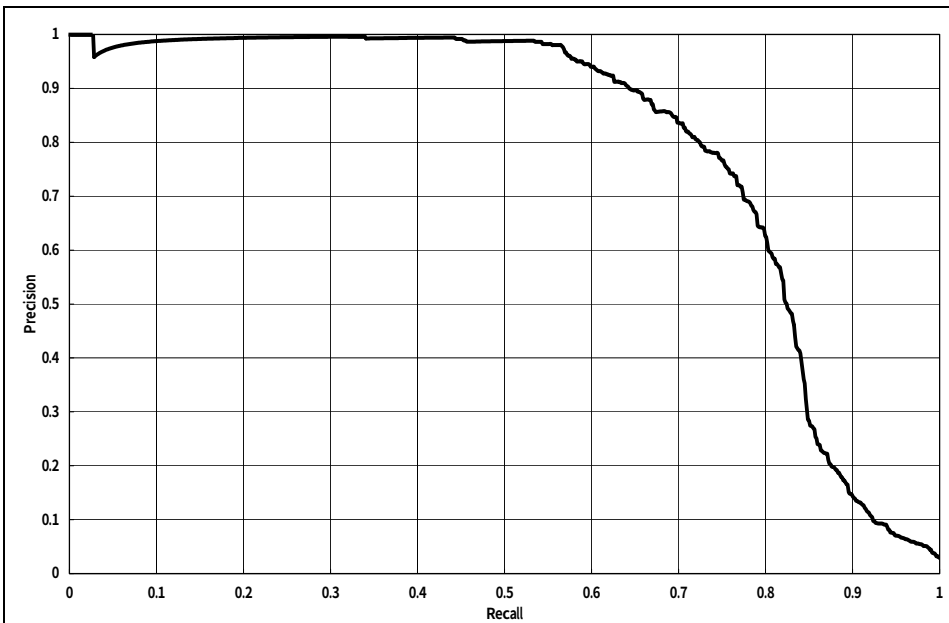

Figure 4.6: Recall Precision Graph for Hydrolases

4.7 Prediction of Hydrolases

This subsection describes the evaluation of the prediction of hydrolases (EC 3.-.-.-). The enzyme activity includes ester hydrolases, glycosylases, aminopeptidases and various hydrolases acting on many kinds of bonds [44, 45].

The recall precision graph of the prediction for hydrolases is shown in Figure 4.6. 22 (2.7%) proteins annotated with hydrolases have been correctly predicted. The 50%, 80%, and 90% of proteins annotated with hydrolases have been predicted with 98.8%, 62.6% and 24.6% precision, respectively. With 80% precision, the 72.5% of proteins annotated with hydrolases have been predicted. The maximum f-measure is 0.765.

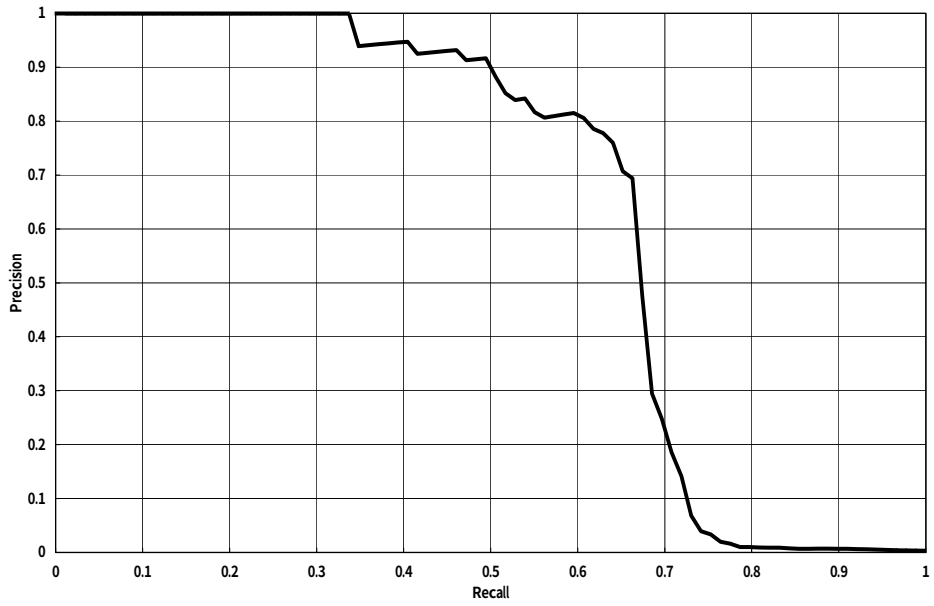

Figure 4.7: Recall Precision Graph for Lyases

4.8 Prediction of Lyases

This subsection describes the evaluation of the prediction of lyases (EC 4.-.-.-). The enzyme activity includes carboxy-lyases, hydro-lyases, ammonia-lyases, and so on [46, 47].

The recall precision graph of the prediction for lyases is shown in Figure 4.7. 30 (33.7%) proteins annotated with lyases have been correctly predicted. The 40%, 60%, and 70% of proteins annotated with lyases have been predicted with 94.6%, 80.6% and

18.6% precision, respectively. With 80% precision, the 60.7% of proteins annotated with lyases have been predicted. The maximum f-measure is 0.696.

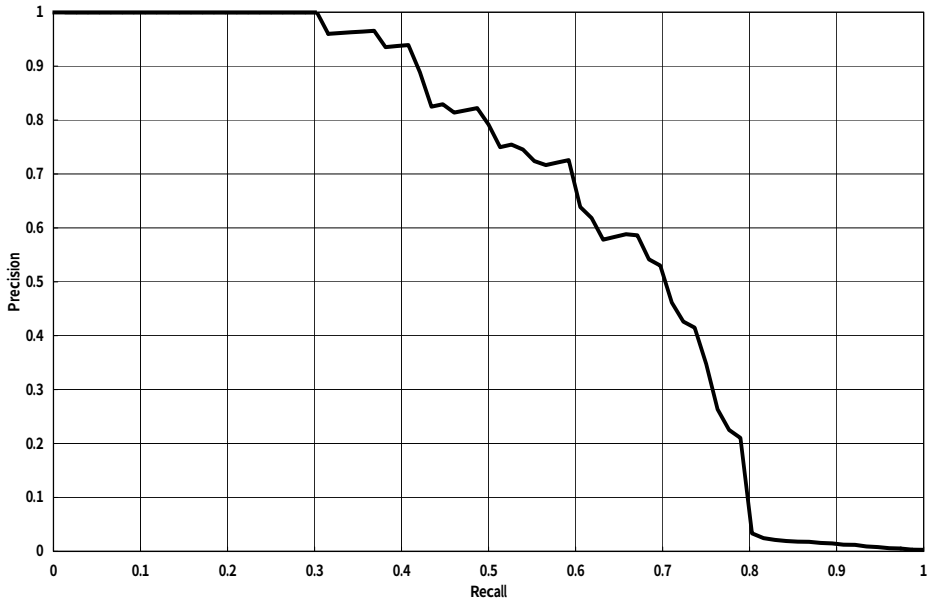

Figure 4.8: Recall Precision Graph for Isomerases

4.9 Prediction of Isomerases

This subsection describes the evaluation of the prediction of isomerases (EC 5.-.-.-). The enzyme activity includes racemases, epimerases, mutases, intramolecular oxidoreductases, Intramolecular lyases, and so on [48, 49].

The recall precision graph of the prediction for isomerases is shown in Figure 4.8. 23 (30.3%) proteins annotated with isomerases have been correctly predicted. The 40%, 60%, and 80% of proteins annotated with isomerases have been predicted with 93.8%, 63.9% and 3.3% precision, respectively. With 80% precision, the 47.8% of proteins annotated with isomerases have been predicted. The maximum f-measure is 0.652.

4.10 Prediction of Ligases

This subsection describes the evaluation of the prediction of ligases (EC 6.-.-.-). The

enzyme activity includes various ligases forming many kinds of bonds [50, 51].

The recall precision graph of the prediction for ligases is shown in Figure 4.9. 11 (7.5%) proteins annotated with ligases have been correctly predicted. The 50%, 70%, and 80% of proteins annotated with ligases have been predicted with 97.3%, 84.4% and 5.7% precision, respectively. With 80% precision, the 47.8% of proteins annotated with ligases have been predicted. The maximum f-measure is 0.770.

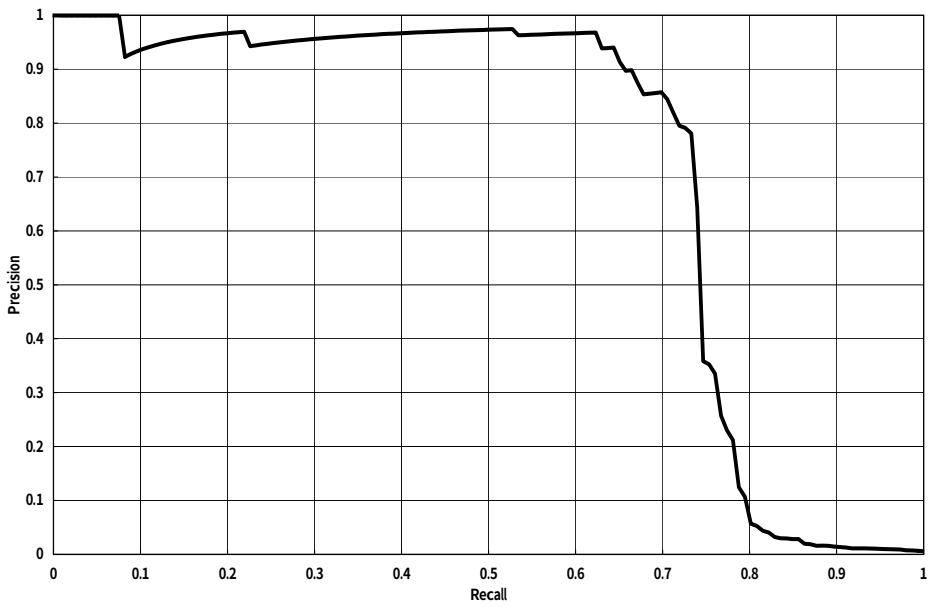

Figure 4.9: Recall Precision Graph for Ligases

4.11 Prediction of Enzyme

This subsection describes the evaluation of the prediction of enzymes. The prediction is regarded as the judgment whether a protein has enzyme activity or not.

The recall precision graph of the prediction is shown in Figure 4.10. 59 (2.6%) proteins annotated with an enzyme have been correctly predicted. The 50%, 70%, and 90% of proteins annotated with an enzyme have been predicted with 96.3%, 80.4% and 30.8% precision, respectively. With 80% precision, the 70.3% of proteins annotated with an enzyme have been predicted. The maximum f-measure is 0.751.

4.12 Conclusion

In the experiments for protein-tyrosine kinase, it scores maximum f-measure of 0.932. The result suggests that the proposed method is quite efficient for a specific enzyme activity.

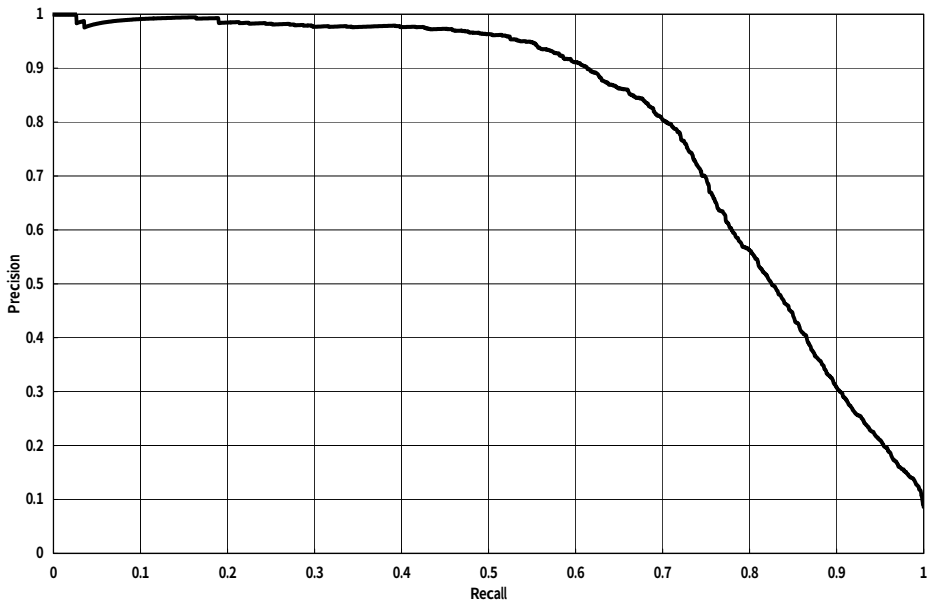

Figure 4.10: Recall Precision Graph for Enzyme

Observing the change of f-measure in a series of experiments for protein-tyrosine kinase, EC2.7.1.-, EC2.7.-.- and transferases, the proposed method is more efficient for a specific enzyme activity than for larger class of enzyme activity. Nevertheless, it scores the maximum f-measure for transferases stays at the high value of 0.786, and the recall precision graph has a good convex upward shape. It suggests that the proposed method is also efficient for predicting a large class of enzyme activities.

The evaluation for large classes of enzyme activities from EC1.-.-.- to EC2.-.-.- clarify that the proposed method is applicable to all classes but there are differences among them. The proposed method is most suitable for predicting transferases. The average of the maximum f-measure from EC1.-.-.- to EC2.-.-.- is 0.717, while the maximum f-measure for whole enzyme activities (EC-.-.-.-) is 0.770. In order to decide whether a protein has enzyme activity or not, the prediction for whole enzyme activities is better than combining the predictions for six large classes. It suggests that

the proposed method is suitable for the prediction of a high level of function. It is one of the advantages of the proposed method.

Generally speaking for these experiments, every recall precision graph has a rectangular shape: holding horizontally at relatively high precision from 0 to a certain value of recall, and slanted to the lower right corner. The difference of the performance among functions mainly depends upon the length of the horizontal part. In a usual information retrieval system, this horizontal holding part becomes quite short when the prediction performance is quite low. The experiments suggest the excellent performance of our method in terms of common sense of information retrieval.

Furthermore, typically in the recall precision graph for isomerase (Figure 4.8) and whole enzymes (Figure 4.10), the decline of precision is slower and the shape is not rectangular but trapezial. In order to predict the Positive Proteins in this region, the prediction method must take account of the subtle similarity among Positive Proteins. It suggests the proposed method has an ability to take account of the subtle similarity sensitively.

5. Prediction of GeneOntology Terms

In this section, the performance of the proposed method is evaluated for several GeneOntology terms. GeneOntology [1, 52] is known as one of the important ontology in biology, and is annotated to proteins. The length of oligopeptide is a parameter of the proposed method and has an impact to the performance of the prediction. In this section, oligopeptide of length 5 is utilised.

5.1 GeneOntology Term and its Annotation to Protein

GeneOntology provides a set of biological terms with some relations among them including part-of relations and is-a relations. GeneOntology terms are divided into three categories, GO component, GO function and GO process. Each GeneOntology term has a unique ID. For instance, 'membrane' is a GO component term whose ID is 'goid 001620'. The version dated 26-Jun-2005 includes of 217,416 GeneOntology terms, which consists of 130,599 GO component terms, 150,641 GO function terms and 143,876 GO process terms.

Proteins in RefSeq are annotated with GeneOntology terms. In RefSeq utilised in this thesis, 4,488 GeneOntology terms are annotated. They consist of 461 GO component terms, 2,124 GO function terms and 1,903 GO process terms.

5.2 Prediction of GO Component

The subsection describes the evaluation of the prediction of several GO component terms including membrane [goid 0016020] and nucleus [goid 0005634]. The numbers of proteins annotated with membrane and nucleus are 2,549 and 4,218, respectively.

5.2.1 Prediction of Membrane

The recall precision graph of prediction for membrane is shown in Figure 5.1. 424 (16.7%) proteins annotated with the GO component term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO component term is predictable with 95.4%, 86.1% and 61.1% precision, respectively. With 80% and 50% precision, the 71.7% and 83.4% of proteins annotated with the GO component

term is predictable, respectively. The maximum f-measure is 0.757.

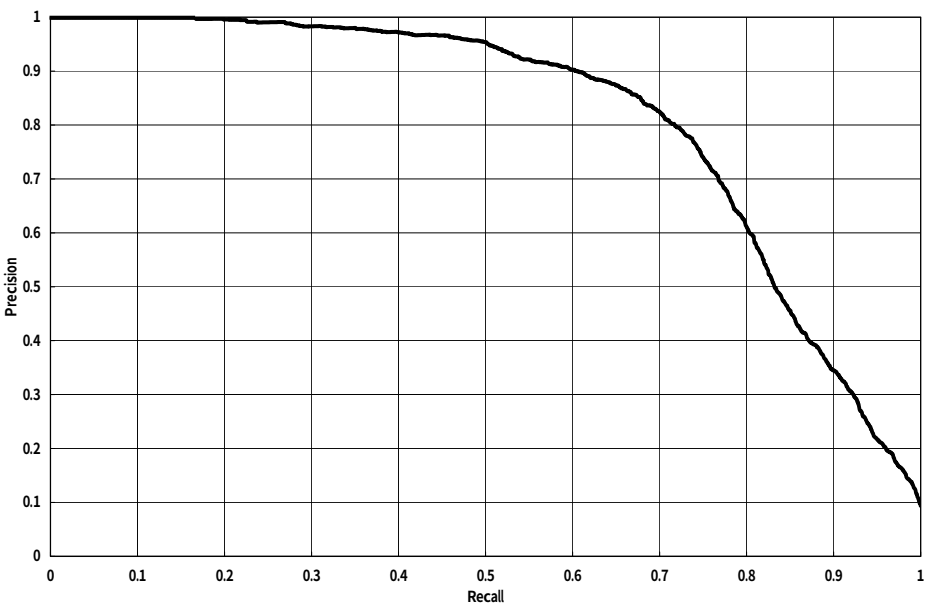

Figure 5.1: Recall Precision Graph for Membrane

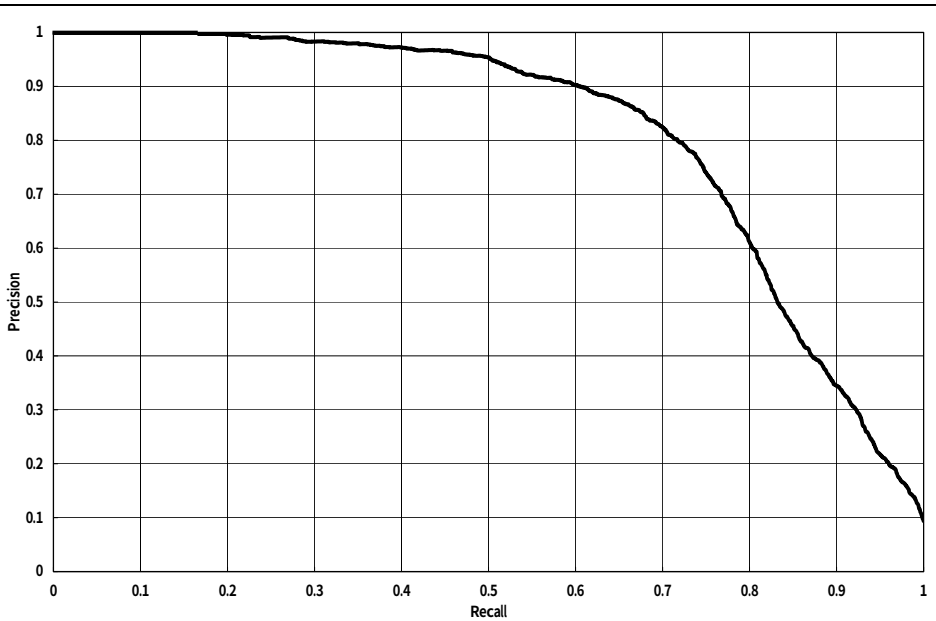

Figure 5.2: Recall Precision Graph for Nucleus

5.2.2 Prediction of Nucleus

The recall precision graph of prediction for nucleus is shown in Figure 5.2. 141 (6.5%) proteins annotated with the GO component term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO component term is predictable with 90.6%, 85.2% and 60.8% precision, respectively. With 80% and 50% precision, the 70.4 and 84.8% of proteins annotated with the GO component term is predictable. The maximum f-measure is 0.753.

5.3 Prediction of GO Function

This subsection describes the evaluation of the prediction of several GO function terms including ATP binding [goid 0005524], hydrolase activity [goid 0016787] and GTP binding [goid 0005525]. The numbers of proteins annotated with ATP binding, hydrolase activity and GTP binding are 1,655, 1,102 and 407, respectively.

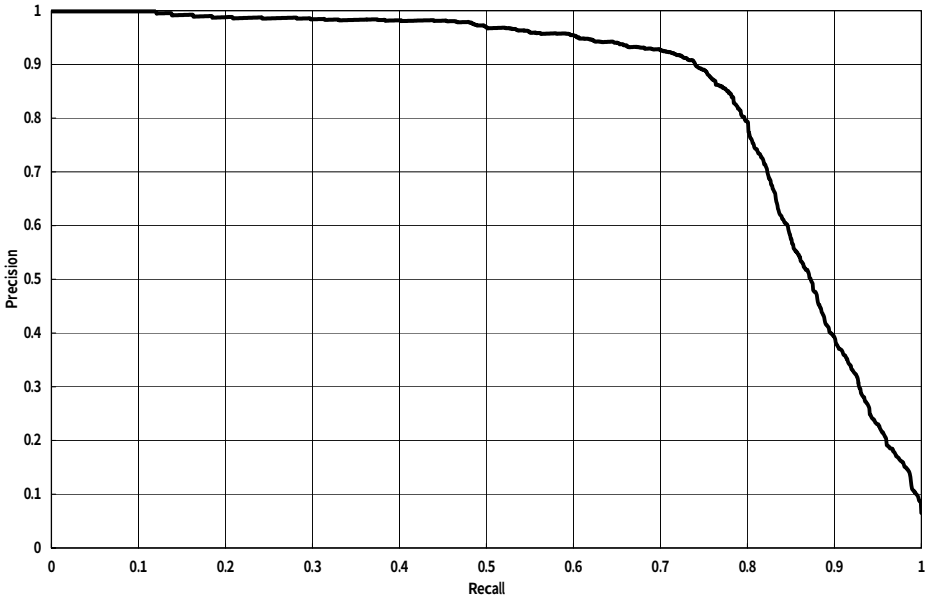

Figure 5.3: Recall Precision Graph for ATP Binding

5.3.1 Prediction of ATP Binding

The recall precision graphs of prediction for ATP binding is shown in Figure 5.3. 199

(12.0%) proteins annotated with the GO function term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO function term is predictable with 97.0%, 95.5% and 79.4% precision, respectively. With 80% and 50% precision, the 80.0% and 62.2% of proteins annotated with the GO function term is predictable, respectively. The maximum f-measure is 0.814.

5.3.2 Prediction of GTP Binding

The recall precision graph of prediction for GTP binding is shown in Figure 5.4. 51 (12.5%) proteins annotated with the GO function term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO function term is predictable with 87.9%, 85.6% and 83.8% precision, respectively. With 80% and 50% precision, the 82.1% and 89.9% of proteins annotated with the GO function term is predictable, respectively. The maximum f-measure is 0.821.

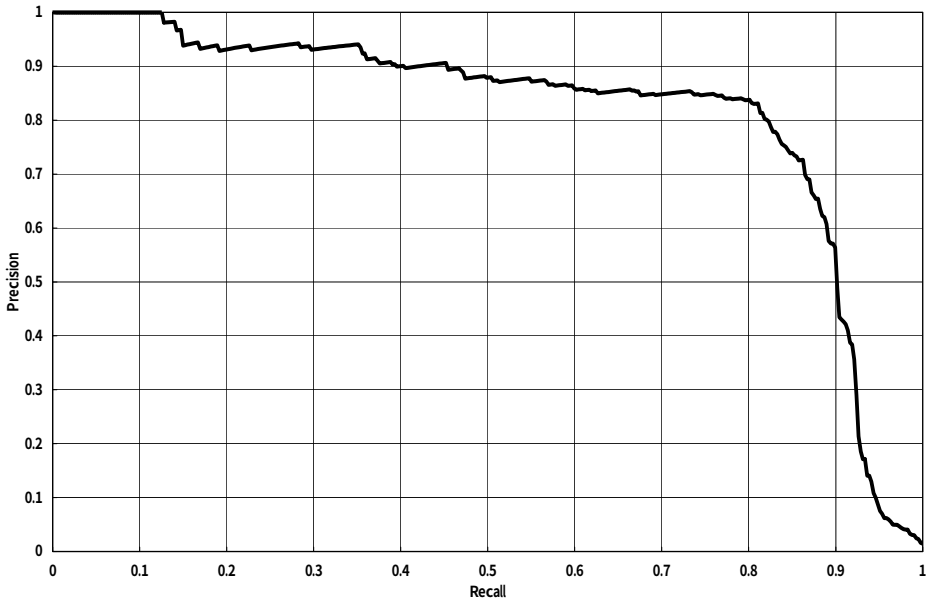

Figure 5.4: Recall Precision Graph for GTP Binding

5.3.3 Prediction of Hydrolase Activity

The recall precision graph of prediction for hydrolase activity is shown in Figure 5.5. 144 (13.1%) proteins annotated with the GO function term are predictable without false

prediction. The 50%, 66%, and 80% of proteins annotated with the GO function term is predictable with 95.5%, 86.3% and 35.0% precision, respectively. With 80% and 50% precision, the 69.6% and 71.1% of proteins annotated with the GO function term is predictable, respectively. The maximum f-measure is 0.751.

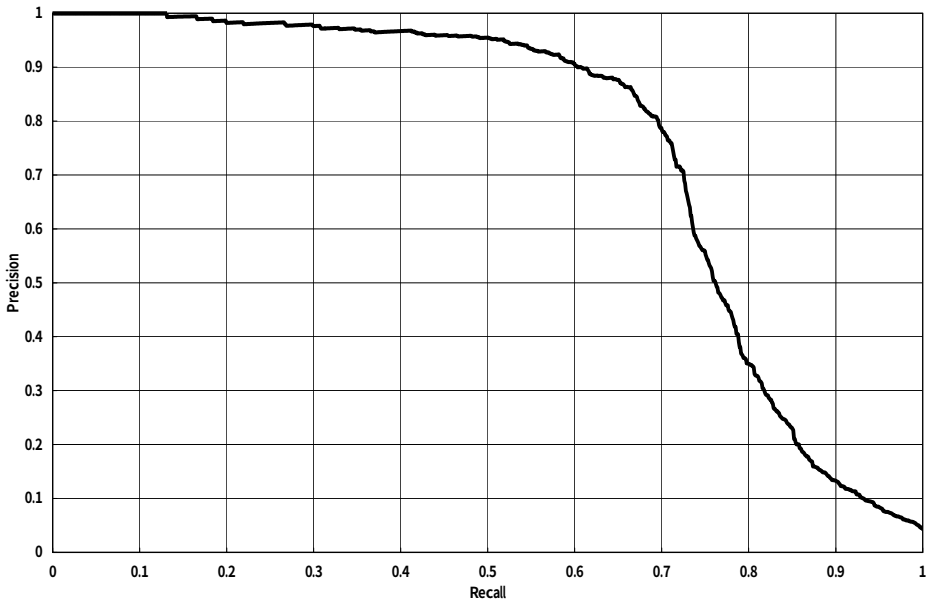

Figure 5.5: Recall Precision Graph for Hydrolase Activity

5.4 Prediction of GO Process

This subsection describes the evaluation of the prediction of several GO process terms including intracellular signaling cascade [goid 0007242] and ubiquitin cycle [goid 0006512]. The numbers of proteins annotated with intracellular signaling cascade and ubiquitin cycle 492 and 334.

5.4.1 Prediction of Intracellular Signaling Cascade

The recall precision graph of prediction for intracellular signaling cascade is shown in Figure 5.6. 43 (8.7%) proteins annotated with the GO process term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO process term is predictable with 95.0%, 87.5% and 63.5% precision, respectively. With 80% and 50% precision, the 73.6% and 81.3% of proteins annotated with the GO

process term is predictable, respectively. The maximum f-measure is 0.769.

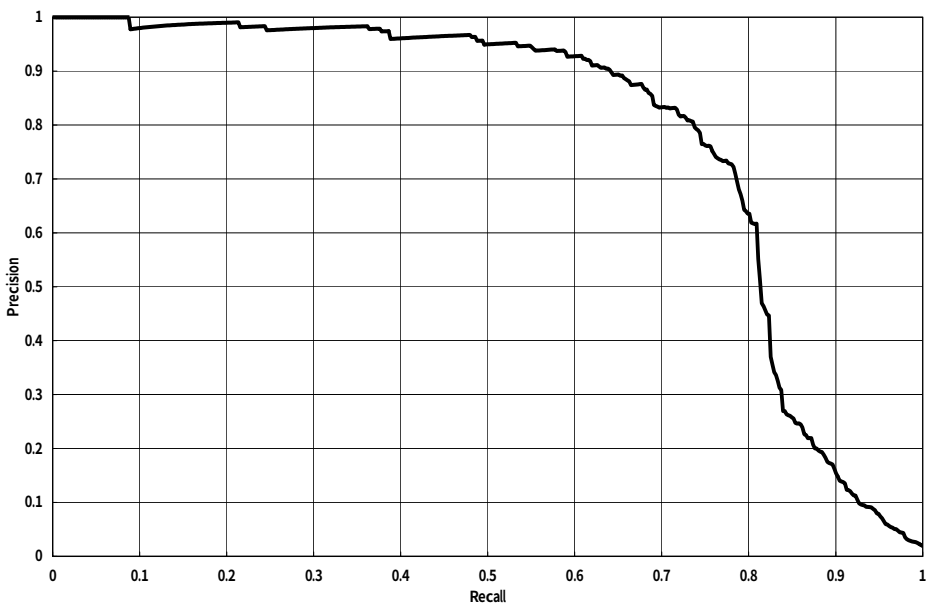

Figure 5.6: Recall Precision Graph for Intracellular Signaling Cascade

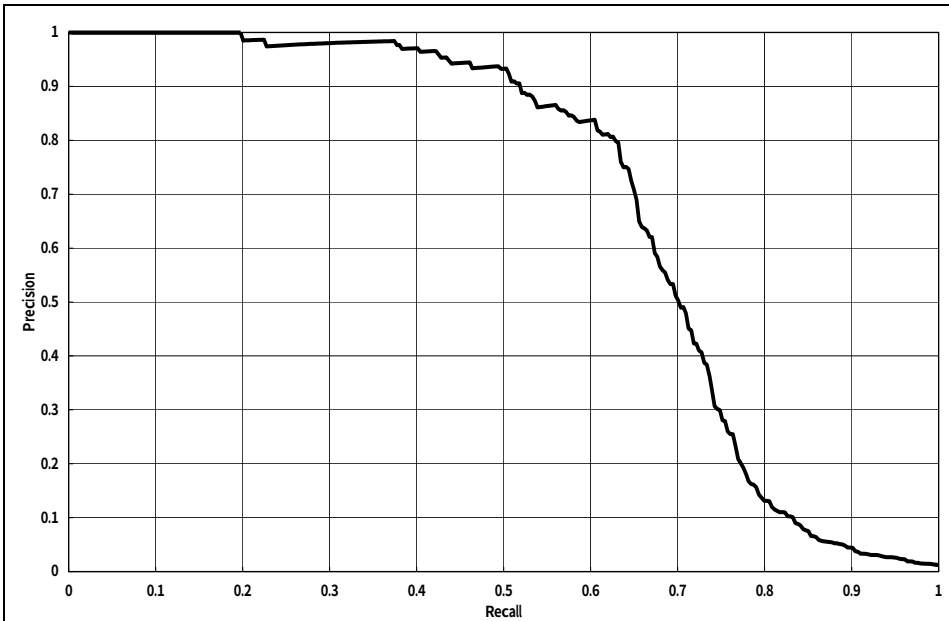

Figure 5.7: Recall Precision Graph for Ubiquitin Cycle

5.4.2 Prediction of Ubiquitin Cycle

The recall precision graph of prediction for ubiquitin cycle is shown in Figure 5.7. 66 (19.8%) proteins annotated with the GO process term are predictable without false prediction. The 50%, 66%, and 80% of proteins annotated with the GO process term is predictable with 95.5%, 85.5% and 35.0% precision, respectively. With 80% and 50% precision, the 62.6% and 76.1% of proteins annotated with the GO process term is predictable, respectively. The maximum f-measure is 0.705.

5.5 Conclusion

In the experiments for GO function terms, ATP binding and GTP binding score over 80% recall with over 80% precision, and the maximum f-measure is greater than 0.8. The results suggest that our method is quite efficient for predicting these GO function terms.

In contrast, the prediction performance for GO process terms is delicate. The results for intracellular signaling cascade are almost equivalently favorable, while ubiquitin cycle scores lower than others (62.6% recall with 80% precision and f-measure = 0.705). We consider that one of the reasons is that the number of annotated proteins is quite less in comparison with cases of other GeneOntology terms.

Generally speaking for all experiments, every recall precision graph has a rectangular shape: holding horizontally at relatively high precision from 0 to a certain value of recall, and slanted to the lower right corner. Prediction performance mainly depends upon the length of the horizontal part. GTP binding (see Figure 11) is an excellent instance. In a usual information retrieval system, this horizontal holding part is so short that the prediction performance is quite low. The experiments suggest the excellent performance of our method in terms of common sense in information retrieval. Furthermore, we can explain the difference between ATP binding and GTP binding by the characteristics of correspondence calculation in our method. Because the correspondence calculation is normalised by the number of included oligopeptides which is proportional to the length, each oligopeptide has larger impact and makes the characteristics of prediction performance more clearly in case of a short protein than a long protein. Because the length of protein annotated with GTP binding (approx. 460 amino acids in average) is quite shorter than ATP binding (approx. 900 amino acids in average), GTP binding results in better than ATP binding.

6. Comparison with Other Prediction Methods

To clarify the performance of the proposed method objectively, this section describes the comparative research with some already proposed prediction methods based on homology search and pattern matching. The length of oligopeptide is a parameter of the proposed method and has an impact to the performance of the prediction. In this section, oligopeptide of length 5 is utilised.

6.1 Homology Search

Homology search refers to scouring a sequence database to find sequences that are likely to be homologous to a given query sequence. BLAST [53, 54, 55], one of the most common homology search tools, is utilised for the comparative studies. BLAST calculates the score of homology, so-called `e-value', for every pair of the query sequence and each sequence in the database, and decides its output using a threshold of e-value. Because the output homologous sequences are ordered by e-value, the most homologous sequence in the databases can be obtained if BLAST finds at least one sequences. In this section, a prediction method of protein functions designed for the comparative studies.

BLAST database is constructed from all Protein Universe. For each protein p of Protein Universe, the most homologous protein except p is searched by BLAST. There are the following cases of the results:

- It results in error. The reasons include that some parameters cannot be calculated in BLAST because the new protein has an error-prone irrelevant sequence. In this case, no functions are predicted because BLAST cannot find any homologous proteins.
- BLAST does not find any homologous proteins. It means that there are no sequences that score higher than the threshold. In this case, no functions are predicted.
- BLAST finds only one homologous protein. In this case the functions of p is predicted as same as ones of the obtained Protein Universe.
- BLAST finds a plural number of homologous proteins. In this case the functions of p is predicted as same as the highest scored one. If there are a

plural number of highest scored homologous proteins, then the predicted functions are a total union of each homologous protein's functions.

6.2 Patten Matching

There are some databases of consensus patterns in the world. In this comparative study, patterns registered in PROSITE [8, 56] are utilised. Each pattern of PROSITE is regarded as a regular expression whose alphabets include 20 symbols of amino acids.

For a new protein, if a consensus pattern is matched to the protein, then the function related to the consensus pattern is predicted to the protein. If there are some patterns for a function, then at least one of the patterns are matched, then it is decided that the function is predicted.

6.3 Comparison Method

While methods based on oligopeptides and homology search can be applicable to any functions fundamentally, a method based on pattern matching is restricted to be applied because pattern matching needs the consensus pattern and there are many functions whose consensus patterns are not found. The comparative studies including pattern matching must be performed for the functions whose consensus patterns are registered in PROSITE. Some of enzyme activities are related to consensus patterns in PROSITE.

Unlike that the proposed method calculate the correspondence of each pair of a protein and a function, the methods based on homology search and pattern matching mentioned in the previous subsections definitely select a set of functions of a given protein. It means that the evaluation of the methods based on homology search and pattern matching does not result in recall precision graph but in a pair of specific values of recall and precision. A comparative study is performed in the following manners:

- The maximum f-measure from the method based on oligopeptides and f-measures from the methods of homology search and/or pattern matching are compared.
- The recall and precision from the method of homology search and/or pattern matching is plotted to the recall precision graph from the method based on oligopeptides. If the plotted point is lower than the curve of the graph, then the method associated with the curve is more effective than the method associated

with the point, and vice versa.

The following subsections describe results of the comparative studies. Table 6.1 shows the number of proteins annotated with each function utilised in the comparative studies. Subsection 6.4 describes a comparative study among the proposed method, homology search and pattern matching. Subsections from 6.5 to 6.9 describe comparative studies between the proposed method and homology search because consensus patterns concerning the functions utilised in these studies have not been found yet.

Table 6.1: Number of Annotated Proteins

Function	Type	Number of annotated proteins
Protein-tyrosine kinase	Enzyme	134 (0.5 %)
Transferases	Enzyme	862 (3.0 %)
Nucleus	GeneOntology	4,625 (16.2 %)
Membrane	GeneOntology	2,588 (9.1 %)
ATP binding	GeneOntology	1,784 (6.3 %)
GTP binding	GeneOntology	425 (1.5 %)

6.4 Comparison for Protein-Tyrosine Kinase

This subsection describes the prediction for protein-tyrosine kinase (EC2.7.1.112) by means of the proposed method, homology search and pattern matching. The prediction for the function by means of the proposed method was already mentioned in the thesis.

At first, the consensus pattern utilised in this comparative study is clarified. Each record of ENZYME Nomenclature Database of Swiss Institute of Bioinformatics [7] includes the corresponding PROSITE document entry accession numbers if exist. For instance, ENZYME Release 37.0 of March 2005 provides the correspondence between protein-tyrosine kinase and PROSITE document PDOC00100.

Each record of PROSITE Documentation File of Swiss Institute of Bioinformatics [8] includes consensus patterns. PDOC00100 in Release 19.4 of 21-Jun-2005 includes three patterns: i) [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K, ii) [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT] (3), and iii) [LIVMFYC]-{A}-[HY]-x-D-[LIVMFY]-[RSTAC]-{D}-x-N-[LIVMFY

C](3).

A consensus pattern of PROSITE is an extended regular expression. A hyphen denotes concatenation. The meaning of each component is as follows:

- N : a single character which denotes the definitive amino acid.
- [LIV] : characters enclosed by square brackets denote the choice of amino acids.
- {PD} : characters (or a single character) enclosed by curly braces denote the exception of amino acids. For instance, {PD} means as same as [ACEFGHIKLMNQ RSTVWY]
- x : it denotes a single occurrence of any amino acid. It means as same as [ACDEF GHIKLMNPQRSTVWY].
- ...(3) : an integer enclosed by parenthesis and attached to any component denotes the iteration of the component, where the integer denotes the number of iteration. For instance, [LIVMFYC](3) denotes three-time iteration L, I, V, M, F, Y or C in any combination, while x(2) denotes any pair of two amino acids.
- ...(5,18) : two integers separated by comma, enclosed by parenthesis and attached to any component denote the iteration of the component, where the pair of integers denotes the range of the number of iteration. For instance, x(5,18) denotes any sequences of any amino acids whose length is between 5 and 18.

Figure 6.1 shows the obtained recall precision graph of the proposed method based on oligopeptide. The results of other methods are superimposed on the graph.

By means of the proposed method, the maximum f-measure is 0.932 as mentioned before. The maximum f-measure is obtained at 93.3% recall and 93.3% precision.

By means of the method based on pattern matching, 666 proteins are predicted and 119 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 88.8 %, 17.9 % and 0.297, respectively.

By means of the method based on homology search, 166 proteins are predicted and 129 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 96.3%, 77.7% and 0.860, respectively.

The performance of pattern matching is quite worse than homology search and the proposed method. The performance of the proposed method is better than homology search. The Proposed method scores f-measure higher than homology search and the point of homology search is located lower than the line of the proposed method in Figure 6.1.

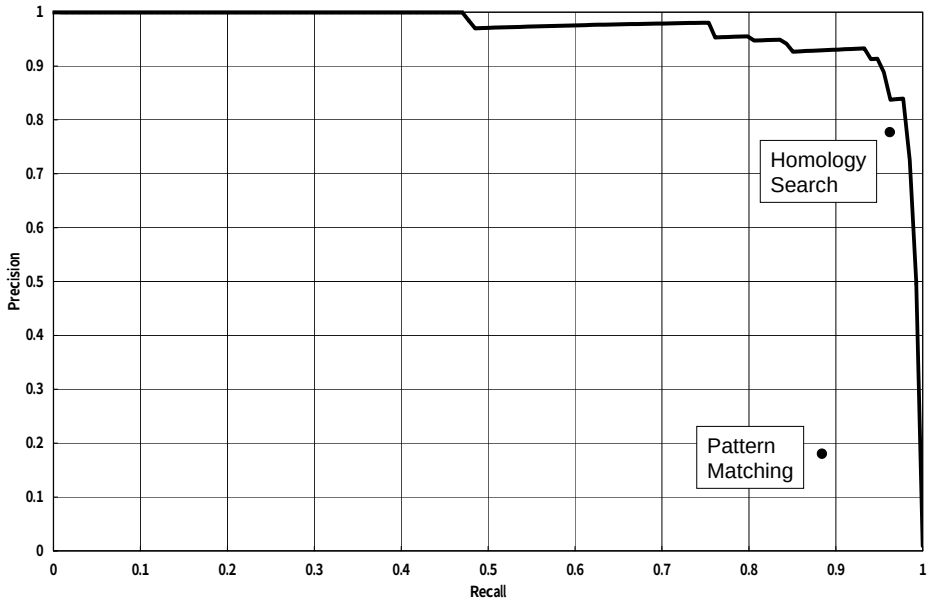

Figure 6.1: Comparison for Protein-Tyrosine Kinase

6.5 Comparison for Transferases

Figure 6.2 shows the obtained recall precision graph of the proposed method based on oligopeptide. The result of homology search is superimposed on the graph. Pattern matching is not applicable to the prediction because the corresponding patterns have not been found.

By the proposed method, the maximum f-measure is 0.786 as mentioned before. The maximum f-measure is obtained at 76.2% recall and 81.2% precision.

By means of the method based on homology search, 1,027 proteins are predicted and 707 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 82.0 %, 68.8 % and 0.749, respectively.

The performance of the proposed method is equal to or better than homology search. The proposed method scores f-measure higher than homology search, while the point of homology search is located upper than the line of the proposed method in Figure 6.2. At the point for maximising f-measure, the precision is better than homology search, while the recall is worse. It suggests that the performance of the proposed method is better than homology search in high precision region for predicting transferases.

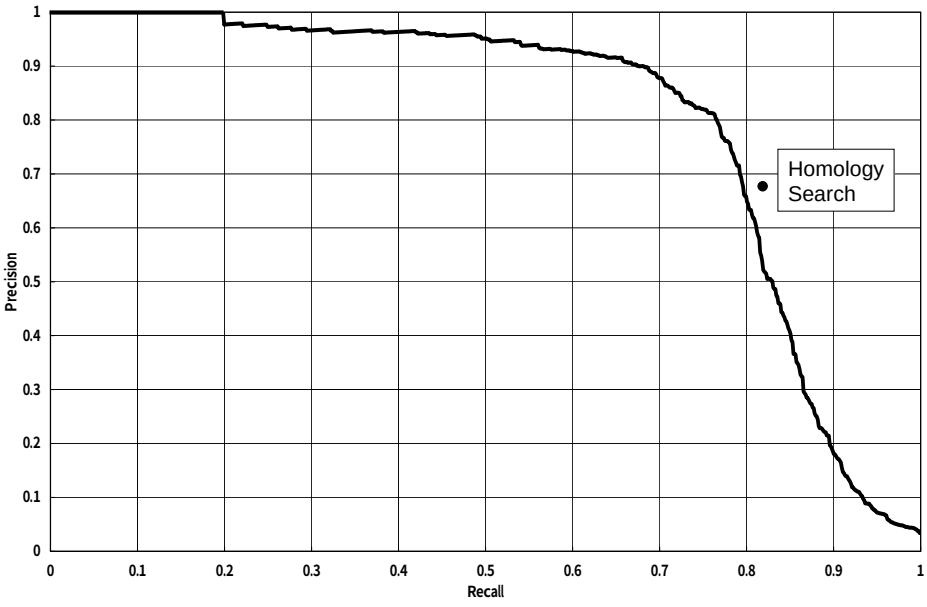

Figure 6.2: Comparison for Transferases

6.6 Comparison for Nucleus

Figure 6.3 shows the obtained recall precision graph of the proposed method based on oligopeptide. The result of homology search is superimposed on the graph. Pattern matching is not applicable to the prediction because the corresponding patterns have not been found.

By the proposed method, the maximum f-measure is 0.753 as mentioned above. The maximum f-measure is obtained at 68.4% recall and 83.6% precision.

By means of the method based on homology search, 5,272 proteins are predicted and 3,407 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 80.7 %, 64.6 % and 0.718, respectively.

The performance of the proposed method is equal to or better than homology search. The proposed method scores f-measure higher than homology search, while the point of homology search is located upper than the line of the proposed method in Figure 6.3. At the point for maximising f-measure, the precision is better than homology search, while the recall is worse. It suggests that the performance of the proposed method is better than homology search in high precision region for predicting the GeneOntology term.

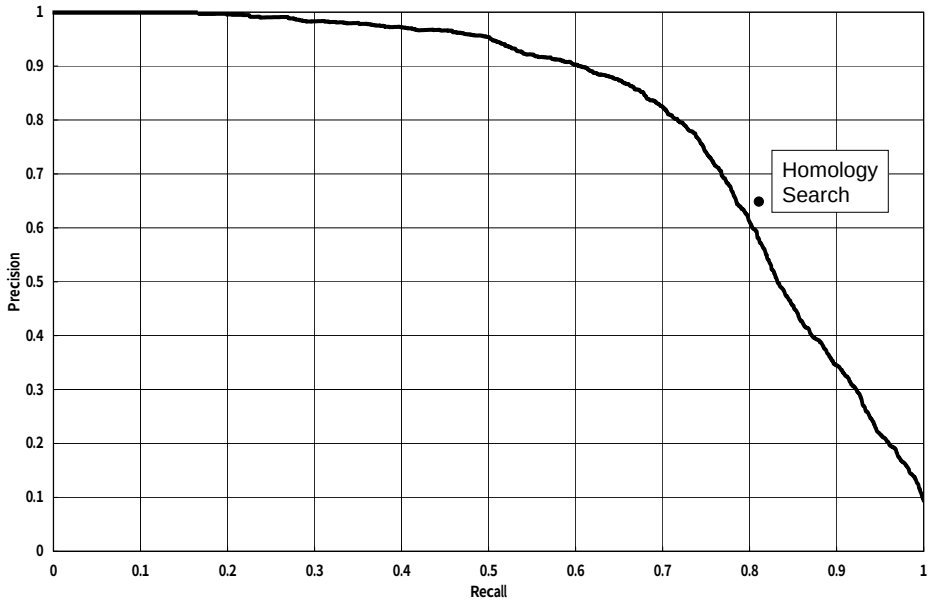

Figure 6.3: Comparison for Nucleus

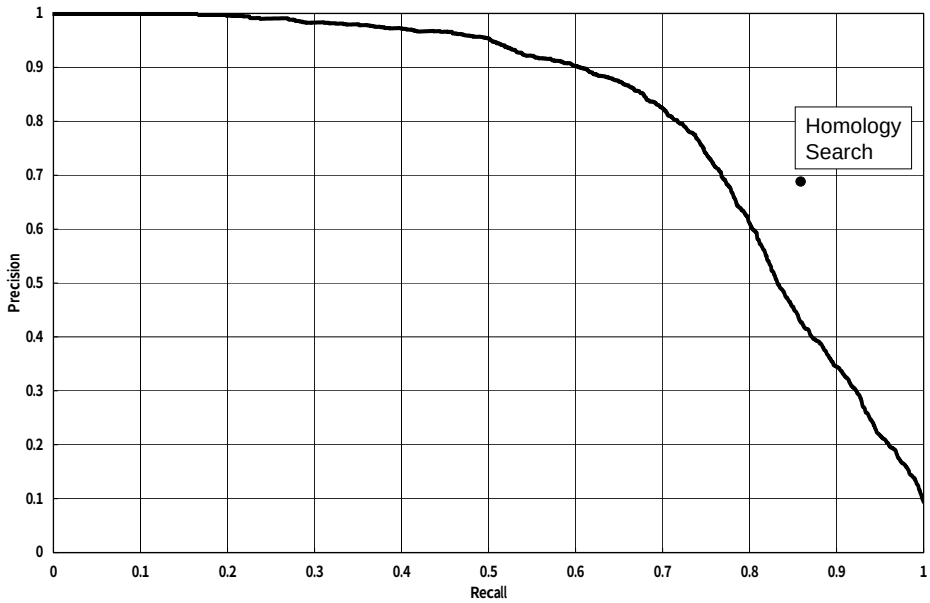

Figure 6.4: Comparison for Membrane

6.7 Comparison for Membrane

Figure 6.4 shows the obtained recall precision graph of the proposed method based on oligopeptide. The result of homology search is superimposed on the graph. Pattern matching is not applicable to the prediction because the corresponding patterns have not been found.

By the proposed method, the maximum f-measure is 0.757 as mentioned above. The maximum f-measure is obtained at 72.3% recall and 79.6% precision.

By means of the method based on homology search, 3,171 proteins are predicted and 2,194 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 86.1 %, 69.1 % and 0.767, respectively.

The performance of the proposed method is worse than homology search. The proposed method scores f-measure lower than homology search and the point of homology search is located upper than the line of the proposed method in Figure 6.4.

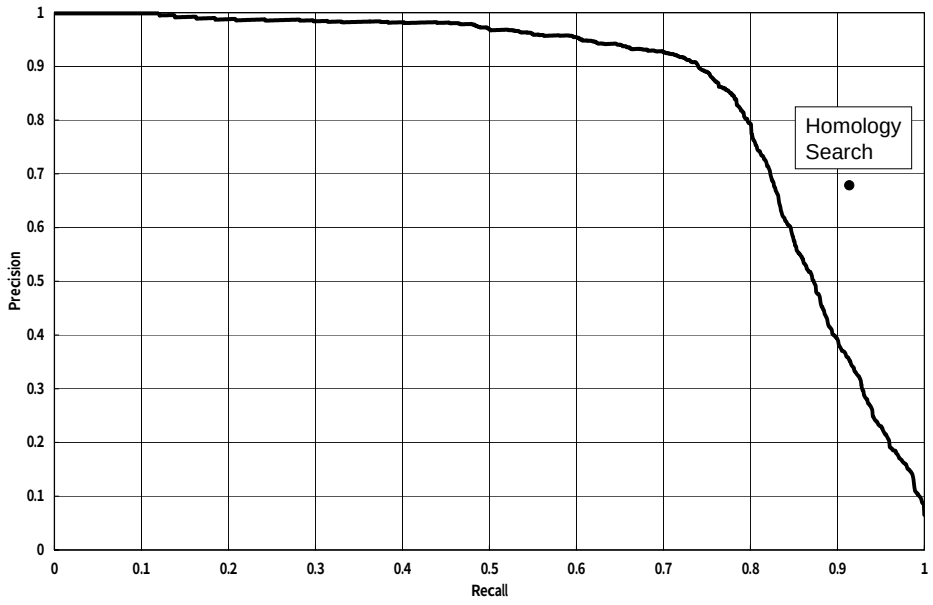

Figure 6.5: Comparison for ATP Binding

6.8 Comparison for ATP Binding

Figure 6.5 shows the obtained recall precision graph of the proposed method based on oligopeptide. The result of homology search is superimposed on the graph. Pattern

matching is not applicable to the prediction because the corresponding patterns have not been found.

By the proposed method, the maximum f-measure is 0.814 as mentioned above. The maximum f-measure is obtained at 75.2% recall and 88.7% precision.

By means of the method based on homology search, 2,217 proteins are predicted and 1,525 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 92.1 %, 68.8 % and 0.788, respectively.

The performance of the proposed method is equal to or better than homology search. The proposed method scores f-measure higher than homology search, while the point of homology search is located upper than the line of the proposed method in Figure 6.5. At the point for maximising f-measure, the precision is better than homology search, while the recall is worse. It suggests that the performance of the proposed method is better than homology search in high precision region for predicting the GeneOntology term.

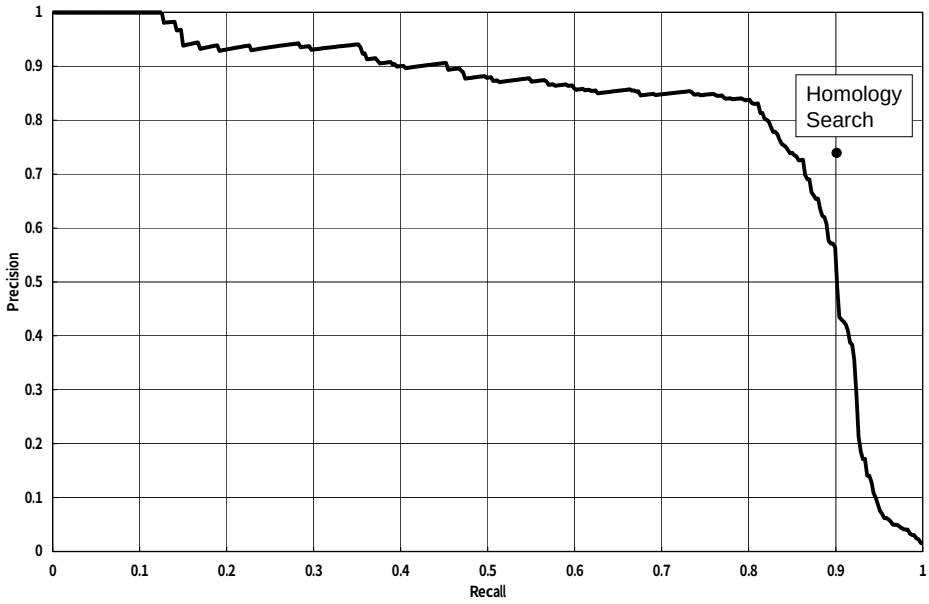

Figure 6.6: Comparison for GTP Binding

6.9 Comparison for GTP Binding

Figure 6.6 shows the obtained recall precision graph of the proposed method based on oligopeptide. The result of homology search is superimposed on the graph. Pattern

matching is not applicable to the prediction because the corresponding patterns have not been found.

By the proposed method, 51 the maximum f-measure is 0.821 as mentioned above. The maximum f-measure is obtained at 81.1% recall and 83.1% precision.

By means of the method based on homology search, 504 proteins are predicted and 368 proteins of them are true positive. Then, the recall, the precision and the f-measure of the method are 90.4 %, 73.0 % and 0.808, respectively.

The performance of the proposed method is equal to or better than homology search. The proposed method scores f-measure higher than homology search, while the point of homology search is located upper than the line of the proposed method in Figure 6.6. At the point for maximising f-measure, the precision is better than homology search, while the recall is worse. It suggests that the performance of the proposed method is better than homology search in high precision region for predicting the GeneOntology term.

6.10 Conclusion

From the results of comparative research with some already proposed prediction methods based on homology search and pattern matching, the performance of the proposed method based on oligopeptide is clarified objectively.

Pattern matching is quite less efficient than the other methods. For instance, in predicting protein-tyrosine kinase, the f-measure by homology search is approximately 2.9 times of the one of pattern matching, and the maximum f-measure by oligopeptides is more than 3 times of the one of pattern matching. The recall by pattern matching is not quite low (88.8 %), while the precision is extremely low (17.9 %). It suggests that the preciseness of pattern, i.e. the strength of regular expression, is quite low.

Table 6.2 summarises the performance of the proposed method and homology search mentioned in this section. Concerning f-measure, the proposed method is better than homology search except for the prediction of membrane. On the other hand, concerning recall precision graph, the point of homology search is located upper than the line of the proposed method except for prediction of protein-tyrosine kinase. It means that oligopeptides method is suitable for protein-tyrosine kinase, while homology search is suitable for membrane. The difference of the length of the focused subsequence is an explicable reason for the result that homology search is more efficient in predicting membrane than the proposed method. Homology search focuses on relatively longer subsequence than the proposed method. On the other hand, a

membrane protein has relatively long typical subsequence in its transmembrane subunits. It is considered that the similarity of the subunits contributes the high performance of homology search in predicting a membrane protein.

Table 6.2: Comparison between Proposed Method and Homology Search

Function	Maximum f-score in proposed method	F-score in homology search	Relation in recall precision graph
Protein-tyrosine kinase	0.932	0.860	OP > HS
Transferases	0.786	0.749	OP < HS
Nucleus	0.753	0.718	OP < HS
Membrane	0.757	0.767	OP < HS
ATP binding	0.814	0.788	OP < HS
GTP binding	0.821	0.808	OP < HS

OP < HS : the point of homology search is located upper than the line of the proposed method.

OP > HS : the point of homology search is located lower than the line of the proposed method.

It was suggested in [2] that the length of sequence have an impact to the performance of the proposed method. Because the voting of each oligopeptide in the longer sequence has relatively smaller influence to *Cor(X, Ps)* than the shorter sequence, it is more difficult to predict a function of the longer sequence more than a function of the shorter one. This is the reason why the performance of predicting GTP binding by the proposed method is better than one of predicting ATP binding [2]. Table 2 shows that this situation is also observed in homology search, i.e. the f-scores for GTP binding and ATP binding is 0.808 and 0.788, respectively. But the relative performance of the proposed method against homology search is not affected by the difference of sequence lengths, i.e. the difference of the maximum f-score in the proposed method and the f-score in homology search for GTP binding ($0.821 - 0.808 = 0.013$) is greater than one for ATP binding ($0.814 - 0.788 = 0.026$). It suggests that the proposed method is more suitable for the prediction of a short protein than homology search.

The comparative studies on the length of oligopeptides suggest that the proposed method is much more efficient than pattern matching and it is more efficient than or equally efficient to homology search on the prediction of protein function.

7. Length of Oligopeptides and Prediction

The length of oligopeptides is an important parameter of the proposed method. The length is thought to make a large impact to the performance of prediction. The longer oligopeptide has relatively high specificity for proteins than the shorter one. The specificity is a basis of the prediction method. In the previous sections, the length of 5 is utilised without any explanation.

To clarify the impact of the length of oligopeptides to the performance of prediction, this section shows the results of comparative research on the length of oligopeptides. In this section, a number of functions are predicted by means of varied lengths of oligopeptides from 1 to 9.

Table 7.1: Length and Variety of Oligopeptides

Length	Variety	Increasing	Unique	Co-occurring
1	20	−	0 (0.00%)	20
2	400	20	0 (0.00%)	400
3	8,000	20	0 (0.00%)	8,000
4	159,724	19.97	479 (0.30%)	159,245
5	2,361,750	14.79	527,213(22.33%)	1,834,537
6	8,049,120	3.41	5,181,054(64.37%)	2,868,066
7	10,138,751	1.26	7,897,565(77.89%)	2,241,186
8	10,483,985	1.03	8,388,139(80.01%)	2,095,846
9	10,589,523	1.01	8,530,459(80.65%)	2,059,164

7.1 Statistics of Oligopeptides

Table 7.1 shows the variety of oligopeptides of the length from 1 to 9 in Protein Universe. Because the variety of amino acids is 20, the theoretical maximum of oligopeptides of length i is 20 raised to the i-th power. In the length from 1 to 3, Protein Universe contains the maximum number of oligopeptides. The increasing rate of the variety descends according the elongation of oligopeptides. Table 1 also shows the increasing rate. In the lengths greater than 6, the variety is almost stable. Some oligopeptides are unique in a specific protein and others are co-occurring in a plural number of proteins. Table 7.1 also shows the breakdown of oligopeptides with the

percentage of unique oligopeptides. All peptides are co-occurring in the lengths less than 5, while the unique peptides are more than half in the lengths greater than 5.

An oligopeptide of length 1 is an amino acid itself. The actual variety of oligopeptides of length 1 counting in Protein Universe is equal to the theoretical one, i.e. 20. The actual variety of oligopeptides shorter than 4 is also equal to the theoretical one. Concerning oligopeptides longer than 3, the variety of longer one is less than the variety of shorter one. Concerning oligopeptides longer than 6, the increase of the variety is saturated. It suggests that the coexistence of oligopeptides more than 6 keeps almost equal characteristics.

All oligopeptides of length 1, 2 and 3 are coexistent to several proteins in Protein Universe and there are no unique oligopeptides for a specific protein. It means that the specificity of longer oligopeptide is less than the one of shorter oligopeptides. Especially concerning oligopeptides longer than 5, the percentage of coexistent oligopeptides are less than one of unique oligopeptides. Concerning oligopeptides longer than 4, the percentage of unique oligopeptides increases along with the length. The ratio of the increase becomes less with the length of oligopeptides. It also supports that the coexistence of oligopeptides more than 6 keeps almost equal characteristics.

From the above observation, it is stated that the oligopeptides of length 1, 2 and 3 are ubiquitous in Whole Proteins. Because of the ubiquity, it is expected that the performance of the proposed method based on oligopeptides of these lengths is quite low. On the other hand, quite long oligopeptides such as ones such as 7, 8 and 9 are maldistributed in Whole Proteins. Because of the maldistribution, it is expected that the applicable functions of the proposed method based on oligopeptides of these lengths is quite low.

Table 7.2: Number of Unique Proteins

Length	1	2	3	4	5	6	7	8	9
Unique Proteins	0	0	0	0	0	0	107	2,900	6,739

For some specific lengths of oligopeptides, there are Whole Proteins each of which does not have any oligopeptides that coexist in the other Whole Proteins. Such proteins are called as Unique Proteins. For each Unique Protein, its functions cannot be predicted from the other protein by the proposed method. Table 7.2 shows the number of Unique Proteins for several lengths of oligopeptides. Unique Proteins are found for oligopeptides longer than 6, and its amount is increased along with the length.

Furthermore, there is a Whole Protein whose length is 7. The method based on oligopeptides more than 7 cannot be applied to the protein.

7.2 Length of Oligopeptides and Transferases

Figure 7.1 shows the obtained recall precision graph of the proposed method based on oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed. Table 7.3 shows the performance of prediction in each length of oligopeptides, i.e. maximum f-measure with the recall and the precision for obtaining the maximum f-measure.

The performance in the lengths of 1 and 2 is quite poor. Especially the performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 5 scores the best of maximum f-measure but the performance in the lengths greater than 4 is quite closed. The convex downward shape of recall precision graph in the length of 3 is different from ones in greater lengths and is commonly observed in informal retrieval research domain. The length of 4 has a similar rectangular shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

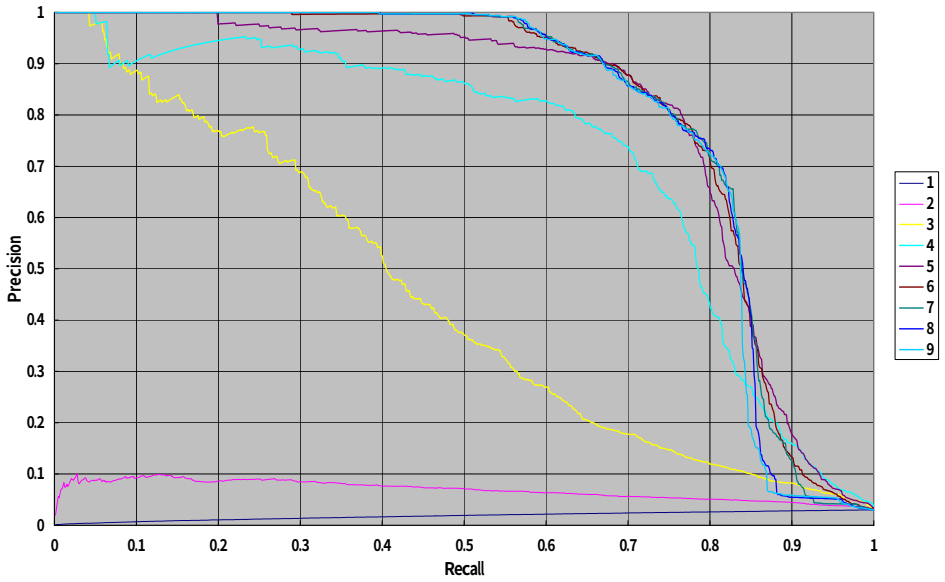

Figure 7.1: Length of Oligopeptide and Transferases

Table 7.3: Maximum F-measure for Transferases

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.059	100.0%	3.0%
2	0.136	26.9%	9.1%
3	0.459	39.8%	54.3%
4	0.719	68.0%	76.4%
5	0.786	76.3%	81.0%
6	0.783	73.9%	83.4%
7	0.782	74.4%	82.5%
8	0.781	74.1%	82.5%
9	0.780	74.9%	81.3%

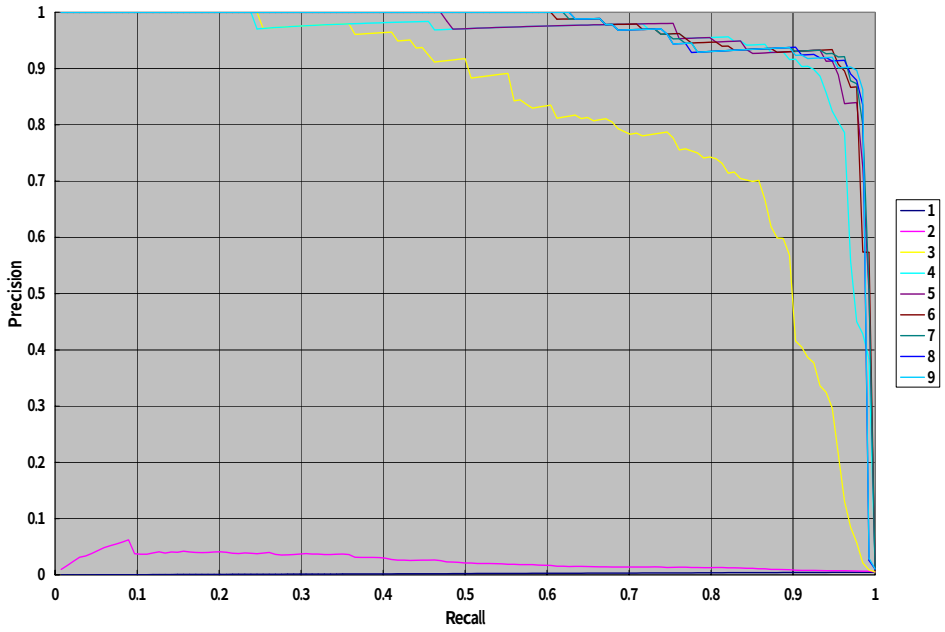

Figure 7.2: Length of Oligopeptide and Protein-Tyrosine Kinase

7.3 Length of Oligopeptides and Protein-Tyrosine Kinase

Figure 7.2 shows the obtained recall precision graph of the proposed method based on oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed. Table 7.4 shows the performance of prediction in each length of oligopeptides.

The performance in the lengths of 1 and 2 is quite poor. Especially the

performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 7 scores the best of maximum f-measure but the performance in the lengths greater than 3 is quite closed. The length of 3 has a similar rectangular shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

Table 7.4: Maximum F-measure for Protein-Tyrosine Kinase

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.009	100.0%	0.5%
2	0.074	9.0%	6.3%
3	0.772	85.8%	70.1%
4	0.912	92.5%	89.9%
5	0.933	93.3%	93.3%
6	0.941	94.8%	93.4%
7	0.942	96.3%	92.1%
8	0.938	96.3%	91.5%
9	0.936	97.7%	89.7%

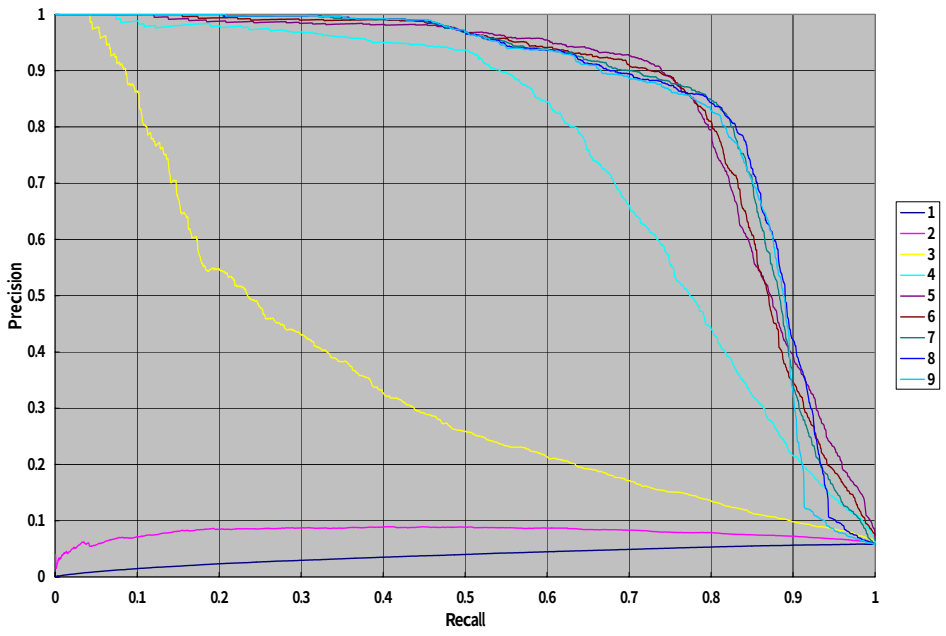

Figure 7.3: Length of Oligopeptide and of ATP Binding

7.4 Length of Oligopeptides and ATP Binding

Figure 7.3 shows the obtained recall precision graph of the proposed method based on oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed. Table 7.5 shows the performance of prediction in each length of oligopeptides.

The performance in the lengths of 1 and 2 is quite poor. Especially the performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 7 scores the best of maximum f-measure but the performance in the lengths greater than 4 is quite closed. The convex downward shape of recall precision graph in the length of 3 is different from ones in greater lengths and is commonly observed in informal retrieval research domain. The length of 4 has a similar rectangular shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

Table 7.5: Maximum F-measure for ATP Binding

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.110	98.2%	5.8%
2	0.152	62.1%	8.7%
3	0.368	35.5%	38.3%
4	0.708	63.7%	79.8%
5	0.814	75.1%	88.9%
6	0.815	75.4%	88.8%
7	0.825	80.7%	84.3%
8	0.823	81.3%	83.4%
9	0.817	79.9%	83.6%

7.5 Length of Oligopeptides and GTP Binding

Figure 7.4 shows the obtained recall precision graph of the proposed method based on oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed. Table 7.6 shows the performance of prediction in each length of oligopeptides.

The performance in the lengths of 1 and 2 is quite poor. Especially the performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 5 scores the best of maximum f-measure but the performance in the lengths greater than 3 is quite closed. The convex downward shape of recall precision graph in the length of 2 is different from ones in greater lengths and is

commonly observed in informal retrieval research domain. The length of 3 has a similar rectangular shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

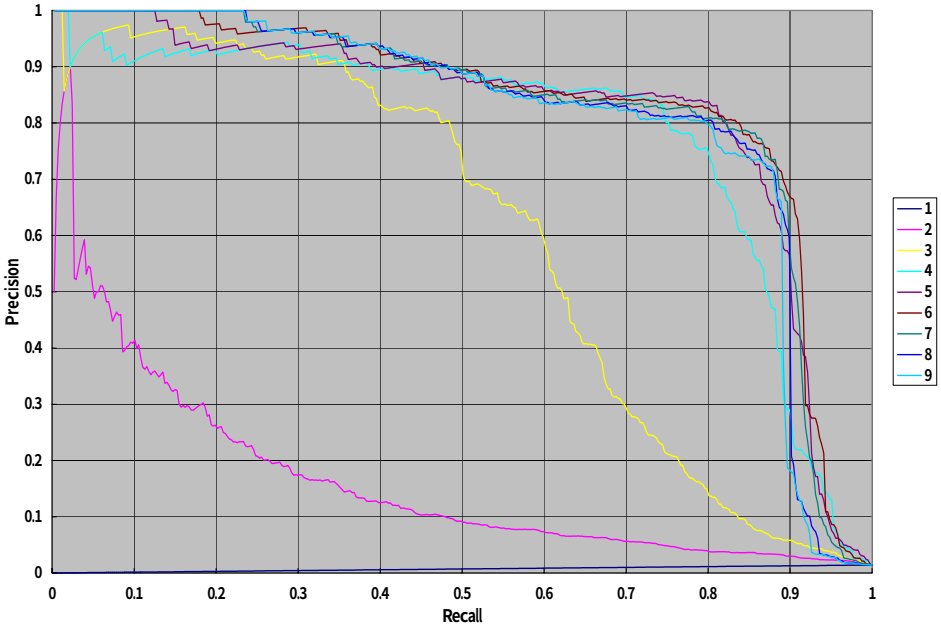

Figure 7.4: Length of Oligopeptide and GTP Binding

Table 7.6: Maximum F-measure for GTP Binding

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.028	100.0%	1.4%
2	0.235	24.3%	22.7%
3	0.610	59.2%	62.9%
4	0.786	74.4%	83.2%
5	0.821	81.1%	83.1%
6	0.819	83.3%	80.5%
7	0.818	85.7%	78.3%
8	0.806	82.8%	78.6%
9	0.800	79.6%	80.4%

7.6 Length of Oligopeptides and Membrane

Figure 7.5 shows the obtained recall precision graph of the proposed method based on

oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed. Table 7.7 shows the performance of prediction in each length of oligopeptides.

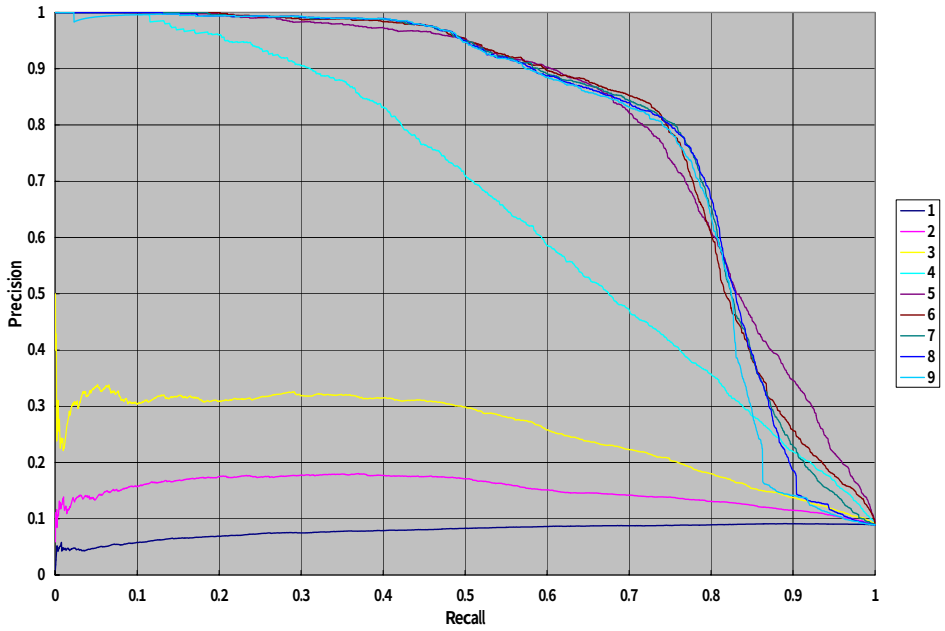

Figure 7.5: Length of Oligopeptide and Membrane

Table 7.7: Maximum F-measure for Membrane

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.166	93.1%	9.1%
2	0.256	48.5%	17.4%
3	0.375	50.8%	29.6%
4	0.601	58.3%	62.0%
5	0.758	72.3%	79.6%
6	0.778	72.5%	83.8%
7	0.778	75.5%	80.2%
8	0.775	74.0%	81.4%
9	0.771	74.0%	80.5%

The performance in the lengths from 1 to 3 is quite poor. Especially the performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 5 scores the best of maximum f-measure but the performance in the lengths greater than 3 is quite closed. The length of 4 has a similar rectangular

shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

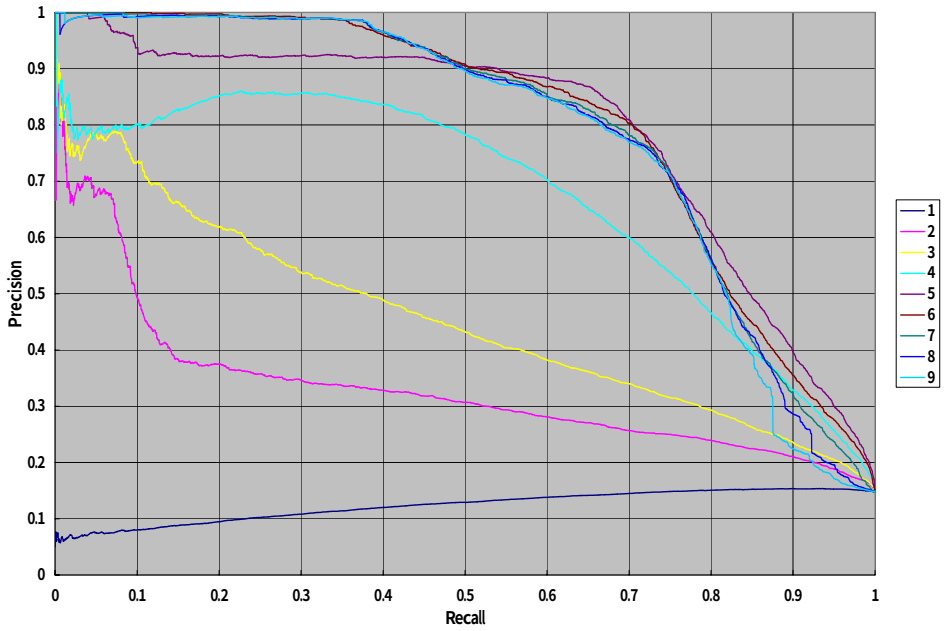

Figure 7.6: Length of Oligopeptide and Nucleus

Table 7.8: Maximum F-measure for Nucleus

Length	Maximum f-measure (*)	Recall for *	Precision for *
1	0.264	96.1%	15.3%
2	0.385	57.1%	29.0%
3	0.470	57.2%	39.9%
4	0.653	63.8%	66.8%
5	0.753	68.6%	83.4%
6	0.750	71.4%	78.9%
7	0.743	72.8%	75.8%
8	0.740	72.2%	75.9%
9	0.737	72.0%	75.5%

7.7 Length of Oligopeptides and Nucleus

Figure 7.6 shows the obtained recall precision graph of the proposed method based on oligopeptide. All graphs for the length of oligopeptides from 1 to 9 are superimposed.

Table 7.8 shows the performance of prediction in each length of oligopeptides.

The performance in the lengths of 1 and 2 is quite poor. Especially the performance in the length of 1 is worse than the theoretical expectation of random sampling. The length of 5 scores the best of maximum f-measure but the performance in the lengths greater than 3 is quite closed. The convex downward shape of recall precision graph in the length of 2 is different from ones in greater lengths and is commonly observed in informal retrieval research domain. The length of 3 has a similar rectangular shape of recall precision graph to ones in the longer oligopeptides but its performance is obviously worse than them.

7.8 Conclusion

Table 7.9 summarises the results of the experiments mentioned in the previous section. For each function, the length which scores the best maximum f-measure and the qualitative evaluation of the shape of recall precision graph are shown. The criteria of qualitative evaluation are as follows:

- −− : Quite poor. The graph is little too high on the right. It means that the performance is worse than the theoretical expectation of random sampling.
- − : Poor. The graph is horizontal line like. The ability of prediction is slightly observed but quite restricted.
- ± : Marginal. The graph is convex downward and commonly observed in information retrieval research domain.
- + : Good. The graph is convex upward or rectangular. But the performance is obviously worse than excellent graphs.
- ++ : Excellent. The graph is rectangular. It is the best length for the prediction or is closed the best.

No predictability is observed in the length of 1. It means that the amino acid profile is not applicable prediction of the functions evaluated in this section. In contrast, oligopeptides is efficient to predict these functions.

The predictability of the oligopeptides whose length is from 2 to 4 depends upon the function to predict. The range of applicable functions of the longer oligopeptides is larger than one of the shorter oligopeptides. For each function, the performance of the longer oligopeptides is better than the shorter oligopeptides.

The oligopeptides whose length is greater than 4 has an excellent predictability.

Figure 7.7 shows the schematic proportion of recall precision graphs for these oligopeptides. A recall precision graph is divided into five regions as follows:

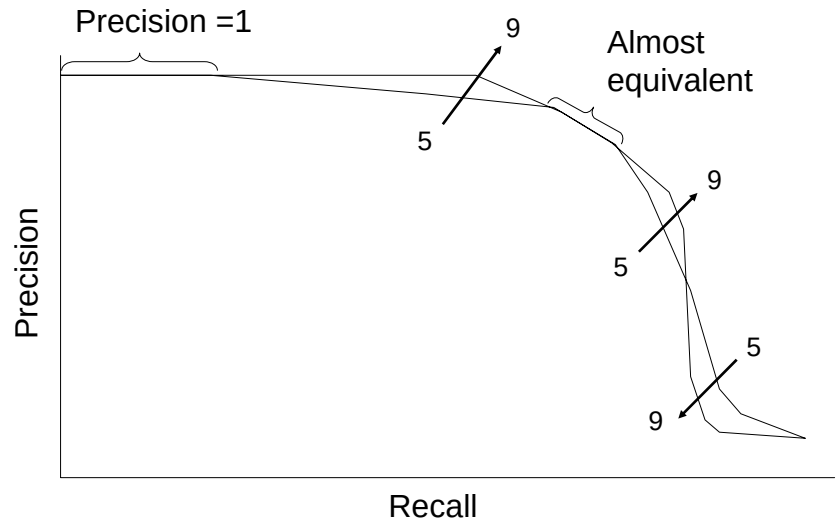

Figure 7.7: Recall Precision Graph of Length 5 to 9

- The horizontal line where precision is equal to 1. The longer oligopeptides keep the length longer. The length of the horizontal line where precision is equal to 1 is positively correlated with the number of Positive Proteins whose *Cor(X, Ps)* is higher than the highest *Cor(X, Ps)* of Negative Proteins. It suggests that the longer oligopeptides avoid a high score of Negative Protein.
- The longer oligopeptides score higher precision than the shorter ones. This is the natural consequence of the difference of the length of the horizontal line mentioned above.
- The graphs meet again. This region is located at the upper right corner of rectangular recall precision graph and the maximum f-measure is obtained from this region. It suggests that the shorter oligopeptides are comparable in the ability of avoiding a high score of Negative Proteins to the longer oligopeptides in the range where the method achieves the best performance to predict.
- The longer oligopeptides score higher precision than the shorter ones again. The reason is same as the relation between the length of oligopeptides and *Cor(X, Ps)* which causes the length of horizontal line where precision is equal to 1, i.e.

the longer oligopeptides avoid a high score of Negative Protein. This region ant the next region mentioned below must be discussed together because these regions are caused by two conflicted relations between the length of oligopeptides and *Cor(X, Ps)*.

- The shorter oligopeptides score higher precision than the shorter ones. In order to predict the Positive Proteins in this region, the prediction method must take account of the subtle similarity among Positive Proteins. The *Cor(X, Ps)* of Positive Proteins in this region by means of the shorter oligopeptides is higher than one by means of the longer oligopeptides. It suggests the shorter oligopeptides amplify a low score of Positive Protein.

This schematic proportion is extended for the oligopeptides whose length is less than 5 when the qualitative evaluation of its recall precision graph is ++ (i.e. excellent). This distinguishing schematic proportion is caused by two conflicted properties: i) the longer oligopeptides avoid a high score of Negative Protein; ii) the shorter oligopeptides amplify a low score of Positive Protein. In other words, the shorter oligopeptides contribute to recall while the longer oligopeptides contribute to precision. It is one of the general findings in information retrieval research domain that recall and precision conflict other.

From the viewpoint of the range of applicable proteins, the longer oligopeptides have a little disadvantage. Table 1 depicts the less co-occurrence of longer oligopeptides. The co-occurrence of an oligopeptide is one of the foundations of the method. If a protein does not share any oligopeptides with any other protein, then the method does not predict any functions for the protein. The number of such inapplicable proteins is positively correlated with the length of oligopeptides. In fact, there are 107, 2,900 and 6,739 inapplicable proteins in Protein Universe for the length of 7, 8 and 9, respectively. There are not inapplicable proteins in Protein Universe for oligopeptides whose length is less than 7. Furthermore, there is a protein whose total length is 7 in Protein Universe. From the viewpoint of efficiency and applicability, it is suggested that five or six is the most acceptable length to predict any functions generally.

In Table 7.9, the performance of prediction of membrane is relatively worse than one of other functions in case of the shorter oligopeptides. The characteristic subsequence of membrane proteins is transmembrane domain. The domain is so relatively long that prediction in case of the shorter oligopeptide is worse. In Table 9, the performance of prediction of GTP binding by means of the shorter oligopeptides is

better than one by means of the longer oligopeptides in comparison with prediction of ATP binding. It also suggested that the characteristic subsequence of GTP binding proteins is shorter than one of ATP binding proteins. Actually, the full sequence of GTP binding proteins is shorter than one of ATP binding proteins and it was suggested in [2] that this is the reason why the performance for GTP binding is better than one for ATP binding. These discussions on the length of a characteristic subsequence and of a full sequence might support each other.

Table 7.9: Summary of Performance in Various Lengths of Oligopeptides

Function	Length for Maximum f-score	Recall precision graph								
		1	2	3	4	5	6	7	8	9
Transferases	5	−−	−	±	+	++	++	++	++	++
Protein-tyrosine kinase	7	−−	−	+	++	++	++	++	++	++
ATP binding	7	−−	−	±	+	++	++	++	++	++
GTP binding	5	−−	±	+	+	++	++	++	++	++
Membrane	7	−−	−	−	+	++	++	++	++	++
Nucleus	5	−−	±	±	+	++	++	++	++	++

−− : Quite poor. − : Poor. ± : Marginal. + : Good. ++ : Excellent.

The thesis investigates the length of oligopeptides and its predictability of some functions including enzymes and GeneOntology. From the viewpoint of the performance and applicability, it suggests that the most acceptable length of oligopeptides is 5 or 6 of generally predicting an arbitrary function.

8. Correspondence between Oligopeptide and Function

The proposed method is based on the correspondence between oligopeptide and function. For each function, the correspondence is calculated from Annotated Protein *Ps* and stored in PepFunc Vector *VEC(Ps)*. This section investigates the correspondence between Protein Universe and protein-tyrosine kinase. This is a typical example of the discussion on the correspondence between oligopeptides and a function. In this section, oligopeptide of length 5 is utilised.

8.1 Statistics on Proteins and Oligopeptides

The number of proteins in Protein Universe is 28,520. 2,361,750 kinds of oligopeptides whose length is 5 are extracted from Protein Universe. The total occurrence of the whole oligopeptides is 14,111,969.

The number of proteins annotated with protein-tyrosine kinase (EC2.7.1.112) is 134 (0.47% of Protein Universe). 58,303 kinds of oligopeptides whose length is 5 are extracted from the Positive Proteins (2.5% of oligopeptides of Protein Universe). The total occurrence of the oligopeptides in the Positive Proteins is 74,989 (0.53% of the total occurrence of oligopeptides in Protein Universe).

8.2 Uniqueness of Oligopeptide and Evaluation Method

The 58,303 oligopeptides of the Positive Proteins contribute the score of prediction *Cor(X, Ps)*. In the evaluation method of the thesis, for each protein in Protein Universe, a prediction is performed by means of the Positive Protein except the protein. If an oligopeptide in the protein is unique in Protein Universe, then it does not contribute the score of any protein in Protein Universe. If an oligopeptide in the protein is unique in Positive Proteins, then it does not contribute the score of any Positive Protein but contributes to the score of Negative Proteins which has the oligopeptide. Consequently, from the viewpoint of the evaluation method, the 58,303 oligopeptides of the Positive Proteins are classified into following 3 groups:

- Unique in Protein Universe. The oligopeptides in this group does not contribute any protein in the course of evaluation. The number of such proteins in Protein Universe is 2,717.

- Unique in the Positive Protein but co-occurring in Protein Universe. It means that there is some Negative Proteins which have such an oligopeptide. An oligopeptide in this group contributes the score of such Negative Proteins only. The number of such oligopeptides is 32,309.
- Co-occurring in Positive Proteins. An oligopeptide in this group contributes the score of proteins which have the oligopeptide. The number of such oligopeptides is 23,277.

For investigating the correspondence between oligopeptides and function by means of the evaluation method, the 23,277 oligopeptides co-occurring in Protein Universe are utilised. The oligopeptides are called as Effective Oligopeptides.

8.3 Oligopeptide of High Correspondence

Table 8.1 shows the classification of Effective Oligopeptides by means of the correspondence for protein-tyrosine kinase. For instance, there are 4,874 Effective oligopeptides whose correspondence for protein-tyrosine kinase is greater than or equal to 1.0 and less than 2.0. The oligopeptides whose correspondence is greater than or equal to 7 are decreased, while there are relatively many oligopeptides whose correspondence is 1, called as Highest Oligopeptides.

Table 8.1: Correspondence of Effective Oligopeptides

Correspondence	<0.1	<0.2	<0.3	<0.4	<0.5	<0.6
Oligopeptides	3,784	4,874	4,144	2,063	1,798	1,959
Correspondence	<0.7	<0.8	<0.9	<1.0	=1.0	
Oligopeptides	1,874	591	431	108	1,647	

Table 8.2: Co-Occurrence of Highest Oligopeptides

Co-occurrence	2	3	4	5	6	7	8	9	10
Oligopeptides	1,043	290	138	41	28	9	21	10	5
Co-occurrence	11	12	13	15	16	17	19	20	>20
Oligopeptides	9	5	13	10	3	5	5	1	11

A Highest Oligopeptide exists only in Positive Proteins. Table 8.2 shows the classification of Highest Oligopeptides by means of the number of the co-occurring

Positive Proteins. For instance, 1,043 Highest Oligopeptides exist in two proteins of Protein Universe and the proteins are Positive Proteins. Oligopeptides with more co-occurring Positive Proteins are regarded to correspond more strongly with the function than ones with less co-occurring Positive Proteins. Table 8.3 lists up the Highest Oligopeptides whose correspondence is greater than 20. For instance, KWMAP exists in 32 Positive Proteins but does not exist in any Negative Proteins. It is suggested that these oligopeptides are specific to and general in the function. Furthermore, both of QPHIQ and PHIQW are Highest Oligopeptides whose occurrence in Positive Proteins is 25. It is predicted that QPHIQW has same property. This prediction is proved by the investigation to oligopeptides whose length is 6. Similarly, SINHTY is Highest Oligopeptides whose occurrence in Positive Proteins is 21. The situation of NVMKI and MKIAD is more complicated. Although VMKIA exists in 21 Positive Proteins, it is not a Highest Oligopeptide because 2 Negative Proteins have the oligopeptide. The investigation to oligopeptides whose length is 6 clarifies that NVMKIA and VMKIAD exist only in 21 Positive Proteins. Furthermore, the investigation concerning the length of 7 shows that NVMKIAD also exists only in 21 Positive Proteins.

Table 8.3: Highest Oligopeptides Co-Occurring in Many Positive Proteins

Co-Occurrence	Oligopeptides
32	KWMAP
29	VKWMA
25	QPHIQ
	PHIQW
	KCIHR
23	WEFPR
	HRIGG
21	SINHT
	NVMKI
	MKIAD
	INHTY

Other than Highest Oligopeptides, oligopeptides whose correspondence with the function and co-occurrence in Positive Proteins are also suggested to be specific to and general in the function. Table 8.4 shows the top 10 oligopeptides in the co-occurrence. For instance, DFGLA exists in 54 Positive Proteins but also in 148 Negative Proteins,

then its correspondence is low in 0.267. In contrast, DVWSF exists in 70 Positive Proteins and its correspondence is high in 0.921 caused by the fact that only 6 Negative Proteins have the oligopeptide. DVWSF is suggested to be one of the most specific to and general in the function. There are three oligopeptides whose length is 6 and include DVWSF as subsequence: ADVWSF, DVWSFG and SDVWSF. It means that DVWSF is always succeeded by a glycine, while it is preceded by an alanine or a serine. Because ADVWSF exists in a Positive Protein and a Negative Protein, it is nor specific to nor general in the function. SDVWSF is in almost same situation as DVWSF.

Table 8.4: Top 10 Oligopeptides in Co-Occurrence

Co-occurrence	Oligopeptide	Correspondence
93	HRDLA	0.861
90	SDVWS	0.804
85	LAARN	0.752
84	DLAAR	0.694
82	RDLAA	0.713
70	VWSFG	0.886
70	DVWSF	0.921
57	IHRDL	0.435
54	DFGLA	0.267
50	WSFGV	0.746

8.4 Conclusion

This section investigates the correspondence between Protein Universe and protein-tyrosine kinase. The oligopeptides whose correspondence is greater than or equal to 7 are decreased, while there are relatively many Highest Oligopeptides. This investigation results in some seemingly important oligopeptides in some reason, because they are specific to and general in protein-tyrosine kinase:

- Highest Oligopeptides with high co-occurrence in Positive Proteins: KWMAP, VKWMA, QPHIQ, PHIQW, KCIHR, WEFPR, HRIGG, SINHT, NVMKI, MKIAD, and INHTY.
- Longer oligopeptides suggested by the above oligopeptides: QPHIQW, SINHTY, NVMKIA, VMKIAD, and NVMKIAD.

- Oligopeptide with high correspondence to the function and high co-occurrence in Positive Proteins: DVWSF.
- Longer oligopeptides suggested by the above oligopeptides: DVWSFG and SDVWSF.

The investigation on each oligopeptide is a great side benefit of a prediction method based on oligopeptide.

9. Conclusion

The prediction of protein function using the sequence is one of the important research topics in bioinformatics. Meanwhile, the statistical characteristics of oligopeptide, relatively short subsequence, have been investigated. This thesis investigates a new method based on oligopeptides. The main purpose of the thesis is to demonstrate that `oligopeptide' enable us to develop an effective method for predicting various protein function. A known function of a protein is regarded to be inherited to its oligopeptides, and the correspondence between oligopeptides and the function is calculated in the whole proteins. In the proposed method, unknown functions of proteins are predicted by means of the correspondence automatically.

The prediction performance of the method is measured for several functions including GO terms and enzyme activities by recall precision graphs using the 28,520 whole human proteins registered in RefSeq. The GO terms include 'membrane', 'nucleus', 'ATP binding', 'hydorolase activity', 'GTP binding', 'intracellular signaling cascade' and 'ubiquitin cycle'. In most cases, it scores 70% recall with 80% precision. The prediction for ATP binding and GTP binding results in quite high performance: it scores 80% recall with 80% precision. Even in the worst case (ubiquitin cycle), it scores 62.6% recall with 80% precision. The enzyme activities include a specific enzyme 'protein-tyrosine kinase' (EC 2.7.1.112) and a large class of enzymes 'transferases' (EC 2.-.-.-). The former and the latter score maximum f-measure of 0.932 and 0.786, respectively. These results suggest that the proposed method is quite efficient for various protein functions.

To clarify the performance of the proposed method objectively, this thesis include the results of comparative research with some already proposed prediction methods based on homology search and pattern matching. For instance, on the prediction of protein-tyrosine kinase, a method based on homology search scores f-measure of 0.860 and a method based on pattern matching scores f-measure of 0.297. These results suggest that the proposed method based on oligopeptides is quite more efficient than pattern matching and is a little more efficient or equal to homology.

The thesis also characterises the relation between the length of oligopeptides and the prediction of protein functions. The performance of prediction is measured for the length of oligopeptides between 1 and 9. For instance, the maximum f-measures for the prediction of transferases are 0.059, 0.136, 0.459, 0.719, 0.786, 0.784, 0.782, 0.781,

and 0.780 using the lengths of oligopeptides from 1 to 9, respectively. These results suggest that oligopeptides of the length of 1 has no predictability, oligopeptides longer than 4 are almost equally effective for all functions. The predictability of oligopeptides of the length between 2 and 4 depends upon the functions, and the longer oligopeptides are more efficient than the shorter ones for each function. The prediction based on oligopeptides utilises coexistence of oligopeptides among proteins. The longer oligopeptides are more versatile than the shorter one because the longer oligopeptide is more varied than the shorter one and the degree of the coexistence is inversely related to the length. Considerations on statistics of oligopeptides results suggest that the most acceptable length of oligopeptides is 5 or 6 of generally predicting an arbitrary function.

The thesis also describes an example of the investigation the correspondence between each oligopeptide and a function, and finds several seemingly important oligopeptides in some reason, because they are specific to and general in a function. This kind of investigation provides the first step to oligopeptide profiling, which will become a useful results of researches based on oligopeptides.

Future works also include improvement of the method and further investigation on every oligopeptide which make predominant impact to predication positively or negatively. The difference of the predictability in various lengths differs according to functions. Future works also include the characterisation, classification and reasoning of proteins from the viewpoint of the difference.

Acknowledgements

First of all, the author gives special thanks to Professor Hirofumi Doi. He gave the author many suggestions, comments and advices continuously from the earliest stage of the research.

The author also thanks Professor Toshio Hakoshima for his comments. The comparative research on the length of oligopeptides was inspired in the discussion with him.

The author also thanks Professor Shigehiko Kanaya for advices for publication of the research results which is mandatory before submitting the doctor thesis. His comments are also helpful to improve the thesis.

The author gives large thanks to Associate Professor Kouichi Doi not only for a huge amounts of discussion with him but also for his efforts on the management of the research laboratory where the research have been carried out.

The author must thank all stuffs and students in Bioinformatics Units in Graduate School of Information Science. Especially, Dr. Tomohiro Mitsumori and Mr. Katsuhiro Iwanishi participate in discussion and help the author concerning the computer facility.

Finally the author thanks to family members, 6 year old Keiko and her mother Minako for their patience. The research was carried out more than 500 km away from home.

References

[1] Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry J., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarkis A., Lewis S., Matese J., Richardson J., Ringwald M., Rubin G., and Sherlock G., Gene Ontology: Tool for the Unification of Biology, *Nature Genetics*, 25:25-29, 2000.

[2] Doi H., Kitajima M., Watanabe I., Kikuchi Y., Matsuzawa F., Aikawa S., Takiguchi K., and Ohno S., Diverse incidences of individual oligopeptides (dipeptidic to hexapeptidic) in proteins of human, bakers' yeast, and Escherichia coli origin registered in the Swiss-Prot data base, *Proceedings of the National Academy of Sciences of the United States of America*, 92(7):2879-2883, 1995.

[3] Pruitt K., Tatusova Y., and Maglott D., NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, 33:D501-D504, .2005.

[4] Salton, G., Automatic text processing - the transformation, analysis, and retrieval of information by computer, Addison Wesley, 1989.

[5] Shug J., Diskin S., Mazzarelli J., Brunk B., and Stoeckert C.: Predicting Gene Ontology Functions from ProDom and CDD Protein Domains, *Genome Research*, 12(4):648-655, 2002.

[6] Zehetner G., OntoBlast Function: From Sequence Similarities Directly to Potential Functional Annotations by Ontology Terms, *Nucleic Acids Research*, 31:3799-3803, .2003.

[7] Bairoch A., The ENZYME database in 2000, *Nucleic Acids Research*, 28:304-305, 2000.

[8] Hulo N., Sigrist C.J.A., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., and Bairoch A., Recent improvements to the

PROSITE database, *Nucleic Acids Research*, 32:134-137, 2004.

[9] Fleischmann, W., Moller, S., Gateau, A., and Apweiler, R., A novel method for automatic functional annotation of proteins, *Bioinformatics*, 15:228-233, 1999.

[10] Henikoff, S. and Henikoff, J.G., Protein family classification based on searching a database of blocks, Genomics, 19:97-107, 1994.

[11] Barutcuoglu Z., Schapire R.E., and Troyanskaya O.G., Hierarchical multi-label prediction of gene function, *Bioinformatics*, 2006 [electric publication version].

[12] Koski L.B., Gray M.W., Lang B.F., and Burger G., AutoFACT: an automatic functional annotation and classification tool, *BMC Bioinformatics*, 16;6:151, 2005.

[13] Almeida, L.G., Paixao, R., Souza, RC., Da Costa, G.C., Barrientos, F.J., Dos Santos, M.T., De Almeida, D.F., and Vasconcelos, AT., A system for automated bacterial (genome) integrated annotation--SABIA, *Bioinformatics*, 20(16):2832-3, 2004.

[14] Ayoubi P., Jin X., Leite S., Liu X., Martajaja J., Abduraham A., Wan Q., Yan W., Misawa E., and Prade R.A., PipeOnline 2.0: automated EST processing and functional data sorting, *Nucleic Acids Research*, 30:4761–4769, 2002.

[15] Wyman S.K., Jansen R.K., and Boore J.L., Automatic annotation of organellar genomes with DOGMA, *Bioinformatics*, 20:3252–3255, 2004.

[16] Andrade M.A., Brown N.P., Leroy C., Hoersch S., de Daruvar A., Reich C., Franchini A., Tamames J., Valencia A., Ouzounis C., and Sander C. Automated genome sequence analysis and annotation, *Bioinformatics*, 15:391–412, 1999.

[17] Abascal F., and Valencia A., Automatic annotation of protein function based on family identification, *Proteins*, 53:683–692, 2003.

[18] Curwen V., Eyras E., Andrews T.D., Clarke L., Mongin E., Searle S.M., and Clamp M., The Ensembl automatic gene annotation system, *Genome Research*,

14:942–950, 2004.

[19] Moller S., Leser U., Fleischmann W., and Apweiler R., EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation, *Bioinformatics*, 15:219–227, 1999.

[20] Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M., and Yuan Y., Predicting function: from genes to genomes and back, *Journal of Molecular Biology*, 283:707-725, 1998.

[21] Bork P., and Koonin E.V., Predicting functions from protein sequences: where are the bottlenecks?, *Nature Genetics*, 18(4):313-8, 1998.

[22] Polacco B.J., and Babbitt P.C., Automated discovery of 3D motifs for protein function annotation, *Bioinformatics*, 2006 [electric publication version].

[23] Levy E.D., Ouzounis C.A., Gilks W.R., and Audit B., Probabilistic annotation of protein sequences based on functional classifications, *BMC Bioinformatics*, 6(1):302, 2005.

[24] Yu G.X., Glass E.M., Karonis N.T., and Maltsev N., Knowledge-based voting algorithm for automated protein functional annotation, *Proteins*, 61(4):907-17,2005.

[25] Watson J.D., Laskowski R.A., and Thornton J.M., Predicting protein function from sequence and structural data, *Current Opinion in Structural Biology.*, 15(3):275-84, 2005.

[26] Chou K.C., and Cai Y.D., Prediction of protein subcellular locations by GO-FunD-PseAA predictor, *Biochemical and Biophysical Research Communications*, 320(4):1236-9, 2004.

[27] Tusnady G.E., Dosztanyi Z., and Simon I., Transmembrane proteins in the Protein Data Bank: identification and classification, *Bioinformatics*, 20(17):2964-72, 2004.

[28] Wren J.D., Hildebrand W.H., Chandrasekaran S., and Melcher U., Markov model recognition and classification of DNA/protein sequences within large text databases, *Bioinformatics*, 21(21):4046-53, 2005.

[29] Saric J., Jensen L.J., Ouzounova R., Rojas I., and Bork P., Extraction of regulatory gene/protein networks from Medline, *Bioinformatics*, 2005 [electric publication version].

[30] McDonald R., and Pereira F., Identifying gene and protein mentions in text using conditional random fields, *BMC Bioinformatics*, 6 Suppl 1:S6, 2005.

[31] Mitsumori T., Fation S., Murata M., Doi K., and Doi H., Gene/protein name recognition based on support vector machine using dictionary as features, *BMC Bioinformatics*, 6 Suppl 1:S8, 2005.

[32] Hofmann O., and Schomburg D., Concept-based annotation of enzyme classes, *Bioinformatics*, 21(9):2059-66, 2005.

[33] http://ncbi.nih.gov/RefSeq/

[34] http://au.expasy.org/enzyme/

[35] Hunter T., Protein kinase classification, Methods Enzymol., 200:3-37, 1991.

[36] Bairoch A., and Claverie J.-M., Sequence patterns in protein kinases, *Nature* 331(6151):22, 1988.

[37] Van Tan, H., Fischer, S. and Fagard, R., Purification of the LSTRA tyrosine protein kinase (p56lck), *European Journal of Biochemistry*, 172(1):67-72, 1988.

[38] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?2.7.1.-

[39] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?2.7.-.-

[40] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?2.-.-.-

[41] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC2/intro.html

[42] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?1.-.-.-

[43] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC1/intro.html

[44] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?3.-.-.-

[45] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC3/intro.html

[46] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?4.-.-.-

[47] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC4/intro.html

[48] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?5.-.-.-

[49] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC5/intro.html

[50] http://au.expasy.org/cgi-bin/get-enzyme-entry-unprecise?6.-.-.-

[51] http://www.chem.qmul.ac.uk/iubmb/enzyme/EC6/intro.html

[52] http://www.geneontology.org/

[53] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., Basic local alignment search tool, *Journal of Molecular Biology*, 215:403-410, 1990.

[54] McGinnis S., and Madden T.L., BLAST: at the core of a powerful and diverse set of sequence analysis tools, *Nucleic Acids Research*, 32:W20-W25, 2004.

[55] http://www.ncbi.nlm.nih.gov/BLAST/

[56] http://au.expasy.org/prosite/

[57] Maurer-Stroh S, Eisenhaber B, Eisenhaber F., N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence, *Journal of*

Molecular Biology, 317(4):541-57, 2002.

List of Publications

Journal Papers

1. Hisayuki Horai, Kouichi Doi, and Hirofumi Doi, Automatic Prediction of Enzyme Activity Based on Oligopeptides. *Journal of Computer Aided Chemistry*, 6:83-89, 2005.

2. Hisayuki Horai, Kouichi Doi, and Hirofumi Doi. A Prediction Method of Protein Function Based on Oligopeptides. (to be submitted)

3. Hisayuki Horai, Kouichi Doi, and Hirofumi Doi. Relation between length of Oligopeptides and Protein Functions. (to be submitted)

International Conference (Refereed)

1. Hisayuki Horai, Kouichi Doi, and Hirofumi Doi (2005). Viewing the Proteome from Oligopeptides and Prediction of Protein Function. *Genome Informatics*, 16(2): 174-182, 2005.