A bioinformatics strategy to produce a cyclically developing project structure

— Comprehensive functional analysis of the drug design target genes —

Makiko Suwa* and Yukiteru Ono

[Translation from *Synthesiology*, Vol.2, No.4, p.299-309 (2009)]

In the midst of the information flood of biological data, the role of the bioinformatics technology rises. This technology is expected to provide information to reduce the risk in the experiments and to help the designing of the experimental protocol. For this purpose, we mainly targeted a G protein coupling receptor (GPCR) and developed a computational pipeline which identifies these genes from genome sequences and performs their functional analyses. The applied results have been worked out into an integrated comprehensive functional analysis database (SEVENS).
This core technology has become the trigger of collaborative researches, which continues today in a spiral evolutionary form. This flow is the dynamic form that continues advancing by the interaction between the research direction determined by three elements as a driving force and the direction of the life science fields progressing rapidly. The three elements are the core technique matured for a long term, the close cooperation with the experiment researcher, and the environment producing technical incubation.

Keywords : G-protein coupled receptor, genome, gene finding, G protein coupling selectivity prediction, a spiral development, SEVENS, GRIFFIN

1 Introduction

Ever since the draft sequence of human genome was published in 2001[1], massive volume of bioinformation began to flood the scene. In about ten years, genome sequence for over 1,000 species of organisms had been decoded. Moreover, with the recent advent of the next-generation sequencer that can decode at a speed that is approximately 1,000 times that of the devices in 2000, there is now a flood of bioinformation. It is certain that an enormous amount of industrially applicable targets (information for genes, RNAs, proteins, etc.) can be obtained in the future, and a highly efficient biochemical experimental technology that can analyze functions will be in demand. However, such analyses require incredible amount of cost and time, and therefore are not feasible at this point.

In this situation, the expectation for bioinformatics technology is increasing. Bioinformatics is a discipline formed by the fusion of biology, information science, and other borderline disciplines. It is a study where large amount of data is processed using a computer, the biological information (code) is digitized and organized as database, the new biological findings are obtained while developing and applying the decoding technology, and the biological phenomena are modeled and described in terms of informatics and physics. It has the advantage of being able to predict and control the behavior of genes that carry biological information. There is the potential that the answer for an analysis that cannot be carried out as a biochemical experiment can be given by the computer at lower cost and higher speed. If this can be accomplished, it may become a navigator that dramatically raises the efficiency of functional analysis experiments.

Among the several industrial application targets, the main biological molecule is the G protein coupled receptor (GPCR)[2]. It exists in the cell membrane, and forms a tubular structure with seven spirals that penetrate the membrane (transmembrane helix). By binding with various ligands, such as neurotransmitters, peptides, odor molecules, and others, from outside the cell, the G protein coupling in the cytoplasm is activated, and the route of information transmission into the cell is determined according to the type (Fig. 1). In many cases, the abnormality of the information transmission system causes severe diseases such as hypertension, cardiac disease, and cancer, and nearly 30 % of the drugs shipped in the world today attempt to control this receptor system. If a drug that can selectively control the activation of G protein is identified, the impact on the market is extremely great. For example, the peptide that controls the expression mechanism of obesity through GPCR is expected to have an enormous market (tens of billions of yen annually) as health food and a useful seed of drugs.

However, biochemical experiments for drug discovery involve extremely high risks and are likened to throwing millions of yen into the sea. For example, the isolation of active peptide with bioactivity is not guaranteed even after years and years of research. Or, in case of searching for the ligand of an orphan receptor whose bonding ligand is unknown, it is necessary to set up a cell environment where

Computational Biology Research Center, AIST 2-4-7 Aomi, Koto-ku 135-0064, Japan ＊E-mail : m-suwa@aist.go.jp

Original manuscript received June 29, 2009, Revisions received October 13, 2009, Accepted October 20, 2009

the GPCR is expressed and can function upon bonding with the G protein. However, since the type of coupled G protein is unknown to the GPCR, it is necessary to set up the experimental systems for all the cell environments where the GPCR is combined with several major types of G proteins. Even if one is capable of reaching this phase, it is difficult to achieve further high efficiency.

In the following chapters, we shall present an approach from the standpoint of bioinformatics to reduce the above-mentioned risks as much as possible, using the GPCR research (hereinafter, referred to as “the Project”) that we have been doing as a model case.

2 Objective of research and research scenario to achieve the objectives

The initial objective of the Project that was started in 2000 was “to present information that contributes to the experimental design by predicting the experimental result with bioinformatics technology, to minimize the risk of biochemical experiment for GPCR drug discovery.”

Specific objectives were: (1) to comprehensively identify and retain the human GPCR genes including the new genes from the genome sequence and to create a database (DB), and to add functional and structural information to these genes in highly efficient manner using the computational method. If the foundation were laid for this DB, it would become easy to find the new GPCRs that may be difficult to isolate or to be expressed by the biochemical experiment.

The other objective was: (2) to develop a program to predict the activation of G protein by entering the ligand and GPCR sequence information and to apply this to orphan receptors whose bonding ligands are unknown. By doing so, the combination of the GPCR and the regulatory drug could be investigated comprehensively, and the design of ligand screening experiment for the orphan receptor would be possible. This would then accelerate the pharmaceutical researches.

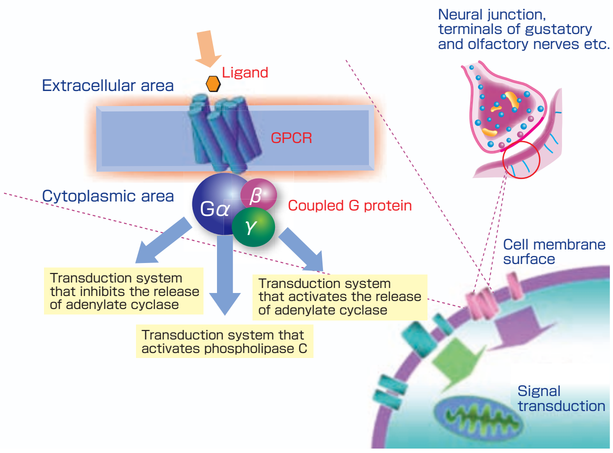
Neural junction, terminals of gustatory and olfactory nerves etc.
Ligand
Extracellular area
GPCR
Cytoplasmic area
Gα
β
γ
Coupled G protein
Cell membrane surface
Transduction system that inhibits the release of adenylate cyclase
Transduction system that activates the release of adenylate cyclase
Transduction system that activates phospholipase C
Signal transduction

Fig. 1 Conceptual diagram for G protein coupled receptors (GPCR).
The GPCRs are present in the cell membrane at places such as the neural junction (right). Various types of molecules (ligands) from extracellular area bond to the structure that is composed by seven transmembrane helices, whereby activating the coupled G proteins, and the signaling pathway to the cell are determined by the G protein type (roughly 3 types) (left).

Above two were the main objectives considered at the start of the Project. The research cycle of bioinformatics, which ranges from basic research to application, is short and as such, the results will instantly become “products” in the form of DB and programs. As the cycle of a typical *Full Research* could be completed in a visible manner, we thought this would be a milestone.

In fact, this would not be the end of the cycle, but it was impossible to correctly draft any future research scenario since the advancement of the life science field was extremely fast. However, we did have some expectations that we would be dealing with a larger flow based on our “product.” Right from the beginning, we could predict that the Project would take several years to be accomplished.

3 First cycle of *Full Research*

Following is a description of the first cycle of the research since the Project was started. It started from identifying the genes from the human genome sequence.

3.1 Gene identification from the genome sequence

The genome is the blueprint of life written on the chromosomes in the cell nucleus. Identifying genes using the computer is like finding a region that has the characteristics of the gene from the DNA (deoxyribonucleic acid) sequence recorded as a long text file. [According to recent understanding, “gene region” has a wide meaning, as it includes the region that codes the functional RNA (ribonucleic acid) as well as the region that codes protein. In this paper, for the sake of discussion, we limit the “gene region” to mean the code region of protein only.]

In most eukaryotes, the genes are separated by several regions called the intron on the genome DNA sequence (Fig. 2). Before this information finally becomes the protein, it is transcribed to mRNA, the introns are cut off, mature mRNA that is bonded only to the exon region of the separated side is formed, and this is then translated into an amino acid sequence. The sequence of three sets of bases that correspond to one code of amino acid during translation is called the codon.

When the DNA sequence is read in order in units of codon, there will be a codon sequence for the starting point. There can be six different codon sequences, including those where one or two bases are shifted from the starting point or where it is read from the opposite end (reading frame). To capture the gene region by the computational method, a model is created by learning the codon at the place where the translation of the protein to amino acid sequence starts (start codon or initiation codon), the codon where it stops (stop codon or termination codon), and the sequence information of the characteristic region such as the boundary between the exon and intron for each reading frame. Then, the regions that match these are extracted.

If the target of search is GPCR, in addition to the general characteristic of the gene, the characteristic region common to protein GPCR is included into the model. These characteristics of the region include the seven transmembrane helixes, as well as the glycosylation site on the NH_2 terminal side of the amino acid sequence, fatty acid binding site of the COOH terminal side, short common functional sequence (functional motif) such as the three amino acids (sequence of Asp, Arg, Tyr (DRY sequence)), and also domains that are globally common over several residues.

The elemental technologies for informatics used in gene identification are groups of programs that capture the above-mentioned characteristics of the gene. An experimental researcher who spends all his/her effort to find new genes without error may be reluctant to use such a program even if the prediction is possible with a certain rate of success. The researcher's demand is that the prediction must be almost entirely correct. Therefore, to allow predictions at extremely high accuracy, we selected a group of appropriate programs from abroad and in Japan, and evaluated their performances.

First we evaluated the program where a known gene sequence is pasted onto the genome by modeling the exon-intron boundaries (ALN[3]), and a program where the expression and transition probability model of nucleic acid base (hidden Markov model) is applied to gene structure (Gene Decoder[4]). We confirmed the maximum length of the gene region from the learned data for nucleic acid sequence region for which the exon-intron structure is decoded in a known gene, and evaluated the ability of the programs to clarify how much upstream and downstream extension from arbitrary exon (additional extension) is needed to cover the entire region of the gene, and studied the sequence resemblance score for identifying the exon most accurately.

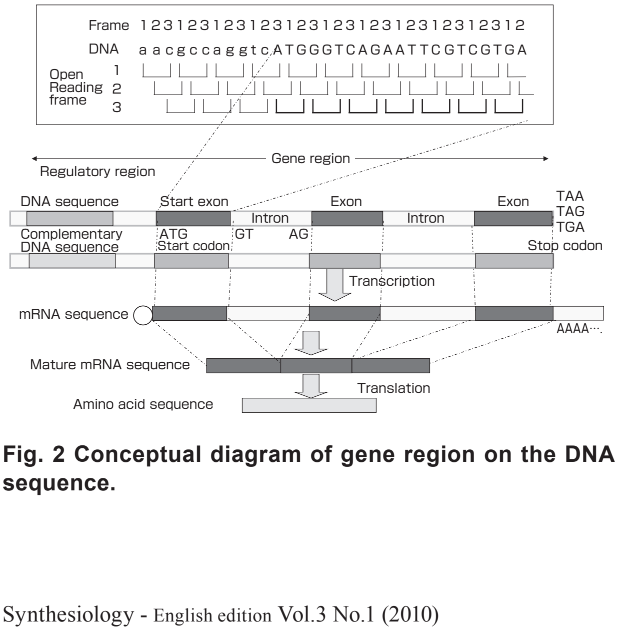

Fig. 2 Conceptual diagram of gene region on the DNA sequence.

Next, as a tool to see whether the gene sequence candidate is actually GPCR or not, the program for sequence investigation (blastp), the program to check the motif characteristic to GPCR (HMMER[5]), and the program to predict the transmembrane helix region (SOSUI[6]) were evaluated. The parameters for selecting GPCR were: resemblance expectation score (E-value) when searching the protein sequence with blastp; E-value for searching the functional motif (Pfam) expressed in the hidden Malkov mode in the HMMER; and the number of predicted helixes in SOSUI. From the learning set including the known GPCR sequence and the non-GPCR sequences in the protein sequence DB (such as UniProt and GPCRDB), the thresholds of the parameters for determining the correct GPCR sequence were set while evaluating sensitivity (percentage of correct predictions among correct sequences) and selectivity (percentage of correct sequences among the predictions). The threshold where almost 100 % selectivity could be achieved while false-negative results (where correct sequence cannot be predicted) were kept to a minimum was defined as "maximum selectivity threshold," while the threshold where nearly 100 % sensitivity could be achieved while false-positive results (where sequence different from GPCR is predicted) were held to the minimum was defined as "maximum sensitivity threshold."

Since the objective was to "understand" the properties of elemental programs that were necessary basic knowledge for solving the issues of the research, this phase could be considered as *Type 1 Basic Research*.

3.2 Gene identification and function analysis pipeline

Based on the research of section 3.1, we developed a system for comprehensively identifying the GPCR gene from the genome sequence. Each elemental program was considered to be a pipe with input and resultant output, and these pipes were joined together step-by-step in optimal order and threshold (SEVENS pipeline, Fig. 3). It is composed of phases for extracting the protein code region from the genome sequence (gene discovery phase), determining the GPCR gene candidate (GPCR gene refinement phase), and adding the function and structure information (functional analysis phase).

This part takes the stance of systematizing by combining the elemental programs and then controlling them as a result and, therefore, may be considered as *Type 2 Basic Research*.

1) Gene discovery phase

The DNA sequence of the genome is scanned for every 6 reading frames, and the corresponding codons are translated to the amino acid sequence, and the fragmented region (corresponding to the exon region) that matches by certain level of resemblance score with the known GPCR amino acid sequence are completely listed (tblasn program). This will narrow down the region where the genes are present, and by using ALN[3], the whole length of the gene corresponding to the known sequence is composed by extending the search region to 1,000 bases upstream and downstream. Also, at the same time, a sequence is obtained by the Gene Decoder[4], which is a probability model of the gene region. Some regions overlap as several sequences match completely or partially, and the parts with significant overlaps are joined to determine the longest amino acid sequence.

2) GPCR gene refinement phase

The determined amino acid sequence is sent to the sequence search program (blastp), the functional motif identification program (HMMER[5]), and the transmembrane helix prediction program (SOSUI[6]) (Fig. 3). By combining the maximum selectivity thresholds and maximum sensitivity thresholds determined for each program in section 3.1, data sets are created from various detection selectivities and sensitivities. While allowing some false-positives (error-in-prediction), if one wished to extract all GPCR, the union of output obtained from the maximum sensitivity threshold for blastp, HMMER, and SOSUI (E-value < 10^{-30}, E-value < 10^{-1}, and predicted number 6~8, respectively) is calculated. This presents 100 % sensitivity at 20.4 % selectivity (level D) for the learning set. On the other hand, the most accurate data set (level A) is the union of output of maximum selectivity threshold of blast and HMMER (E-value < 10^{-80}, E-value < 10^{-10}, respectively). This shows 99.4 % sensitivity and 96.6 % selectivity for the learning set. Also, we created level B (sensitivity 99.8 %, selectivity 70 %) and level C (sensitivity 99.9 %, selectivity 48.4 %) data sets as intermediates between the two levels. Finally the dataset is matched with sequence data for non-GPCR genes, and the wrongly predicted sequences are eliminated.

3) Functional analysis phase

Using the identified GPCR sequence, the sequences related to E-value < 10^{-30} are grouped together, and added to the known family. The sequences that show resemblance of 96 % or over at 100 residues or more for the known GPCR sequence are considered the same as the known sequence, and any other sequences are considered new sequences. If stop-start codons are found in the exon region, it is considered a pseudo-gene. Based on the analysis conducted in the GPCR gene refinement phase, the functional and structural information such as the coordinates on the chromosome, the number of exons, the length of sequence, the sequence search information, the transmembrane helix region, the functional motif region, and the domain region are added to each sequence.

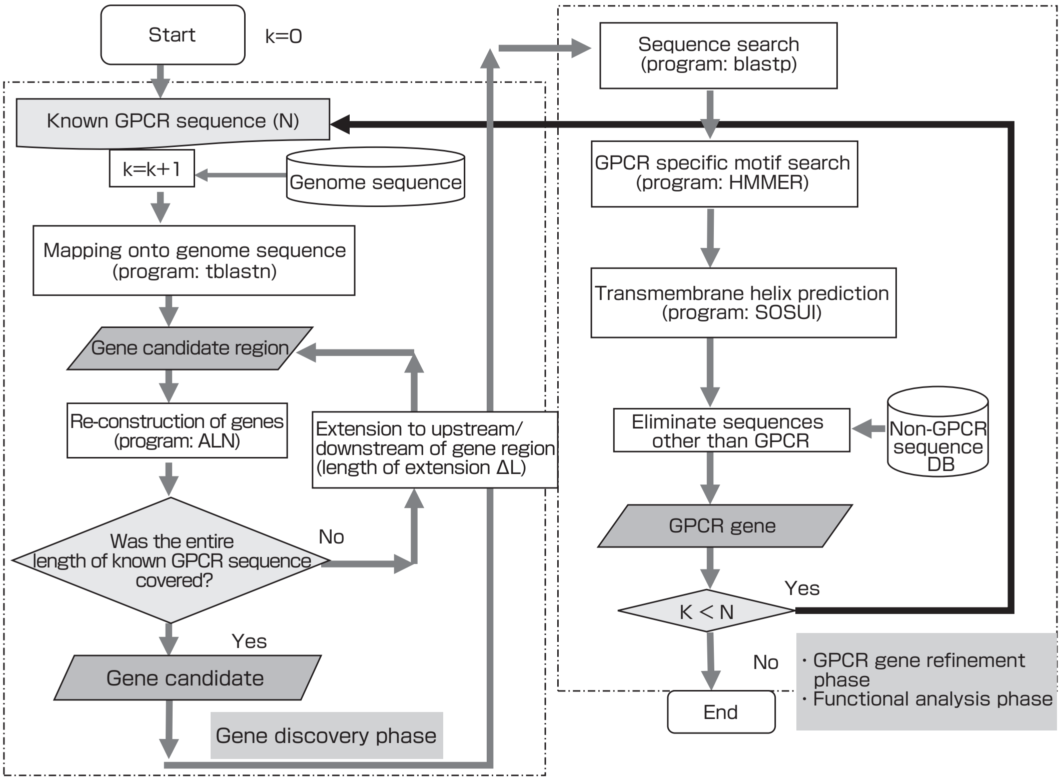

Fig. 3 SEVENS pipeline.

This is an analysis pipeline where various tools are combined sequentially at optimal threshold and order, to comprehensively identify the GPCR genes from the genome sequence.

3.3 Output of the Project

When all GPCRs were identified from the human genome, we obtained 827 level A, 1300 level B, 1517 level C, and 2109 level D sequences. While there were higher possibilities of false-positives (where wrong sequences may be predicted) in sets with higher numbers, they also had higher chances of including new GPCRs. Interestingly, it was found that the majority of the GPCRs were concentrated on chromosome 11, they were dominated by olfactory receptors, and the chemokine receptors were concentrated on chromosome 3. This finding was possible for the first time through this comprehensive gene identification. We applied for patent in 2002 for several hundred sequences that were determined to be new sequences. A certain pharmaceutical company requested disclosure and we received income from this disclosure. Hence, we were able to produce results of the *Product Realization Research*.

The GPCR sequence with additional structural and functional information using the computational method was organized as a database and publicized in 2003 (SEVENS[7] http://sevens.cbrc.jp/1.20/, the very first version). At this point, the core technology was completed to a point, and the first cycle of the Project that started from scratch came to a milestone.

4 Project that undergoes cyclic development

4.1 Hop: Core technology development of the whole Project

The Project that started in 2000 completed one cycle consisting of analysis of elemental technology, systemization, and product realization, and still continues after publication on the website. If the "first cycle of *Full Research*" as described in the previous section was the "hop" of the triple jump, the leap continued in "step" and "jump." Following is the description of the development into joint research, and further technological development through joint research.

4.2 Step: Feedback to core technology through collaboration with industry and academia

In 2002, we experimentally confirmed the expression of several sequences in human tissue for the new GPCR in SEVENS jointly with companies, and applied for patent for particularly important sequences. The fact that the expressions were confirmed for genes predicted by the computational method demonstrated the adequacy of our policy.

However, we also had issues. As a method for confirming the gene expression, we used the polymerase chain reaction (PCR) where minute nucleic acid sequence samples could be rapidly multiplied in a short time. However, it is desirable that the sequence used in PCR analysis have full length with the correct terminals on both ends. However, we found that there were many cases where the terminals were lacking as a result of failure of start (or stop) exon identification in the predicted genes. Most of these were long genes composed of several exons, and since they extended into a wide region, the parameter set in section 3.1 (1,000 bases) was not sufficient as the parameter for the extension of the gene region. Therefore, we investigated the gene existence region to cover a wider area than the commonly assumed area, and it was surprisingly found that it was necessary to extend an arbitrary exon upstream and downstream to 140,000 bases.

Although the subject of SEVENS pipeline was GPCR, it is applicable to other types of protein if the parameters at each phase are changed. We attempted this in the joint research with a venture research center of the University of Tokyo in 2002. In chronic inflammatory diseases, such as rheumatoid arthritis and multiple sclerosis, the immunocytes aggregate excessively at the site of the chronic inflammation and destroy the tissue. This is because the migration of the immunocytes is triggered when the protein called chemokine binds with its GPCR (CCR2). There was a competition for the search of a molecule that inhibits the binding of chemokine (antagonist). However, to avoid the side effects that were expected to occur when the antagonist intercepted the chemokine receptors of different subtypes with structures similar to CCR2 that may be active during organ formation, cell multiplication and differentiation, there was demand to look for a molecule that controlled CCR2 through a different route from the antagonist.

The experimental research showed that a new gene that gathered specifically at the C-terminal of the CCR2 (FROUNT) could be the candidate. We found that this was a long protein composed of 600 residues, where multiple helixes appeared repeatedly. Also, as a result of searching the genome for the characteristic of having several short and weak motifs, we found that there were only two regions that completely matched the new gene, but there were several that showed matches at a weak score. This study was published in *Nature Immunology*[8]. The technologies that were re-investigated in the two joint researches were reflected in the SEVENS pipeline.

4.3 Jump: The development of the new function prediction program

Joint research with a pharmaceutical company started in 2004. Here, a computation system to efficiently and comprehensively screen the ligands that regulate the activation of G protein selectively was build, and we applied this to the ligand screening of the orphan receptors whose binding ligands were unknown.

First, we selected 108 novel human GPCRs from the level A data set of SEVENS. These were orphan receptors. Next, the ligands to be used in screening were comprehensively identified from the human genome based on known peptide ligands after optimizing the gene identification pipeline for peptide ligand search.

On the other hand, we developed a program to monitor the G protein activation. First, using the sequence with known coupled G protein and binding ligand ($G_{i/o}$ type: 61, $G_{q/11}$ type: 47, G_s type: 23), we determined the parameter effective for determining and categorizing the types of coupled G protein and the optimal determination plane from the physicochemical parameters of various sites of the ligand, GPCR, and G protein using the support vector machine (SVM) method that is thought to have the highest identification performance. Using the optimized parameter[9] and determination plane, we created a hierarchical determination program (GRIFFIN) that conducted binary determination of $G_{i/o}$ or $G_{q/11}$ from the remnant after selecting the G_s bonding type upon entering the ligand molecular weight and GPCR. The prediction could be conducted at sensitivity and selectivity of 85 % or higher[10].

Using the above, it is possible to predict the G protein type that is activated in the downstream of signal transmission by the receptor to which certain peptide ligand bonded, based on the database of the ligand that bonds with GPCR, and this will be useful in designing the evaluation system for the expression of the receptor. GRIFFIN would be used for predicting the GPCR with unknown functions in the functional analysis phase of SEVENS.

4.4 Hop again: Type 1 Basic Research to up-scale the research

Up to this point, the research handled human genome only, but this research, in principle, is applicable to genomes of other species. From 2005, we participated in the Scientific Research on Priority Area of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and started comparative genome research in earnest. It was necessary to alter the SEVENS pipeline for other organisms. Based on the genome sequences of about dozen eukaryotes and over 200 prokaryotes that were available at the time, we studied the resemblance expectation score (E-value) when mapping the known gene onto the genome sequence, and the additional extension upstream and downstream to the gene candidate regions. Using the improved pipeline, while GPCR could not be identified from prokaryotes, in eukaryotes, we found some GPCR in yeasts, a dozen in plants, about 200 in insects, several hundreds in fish and birds, and several hundreds to thousands in mammals. Among insects, nematodes, and vertebrates, the minimal number of receptors necessary for life activities such as neurotransmission and intercellular interaction were retained in all organisms, and the types of receptors for complex functions increased dramatically in vertebrates. The receptors for chemical substances in the environment were distributed uniquely in different organisms according to atmospheric or aqueous environments. For example, mammals had high percentage of olfactory receptors in the GPCR genes, and they dominated about 70 %. This indicates that they increased rapidly with repeated high-density gene duplication[11]. The SEVENS pipeline for multiple organisms was almost completely automated at this point, and it became possible to continue analysis even with increased number of organisms.

4.5 Step again: use of pipeline with new protocol

We received high acclaim for identifying and publicizing the GPCRs of various organisms, and from 2007, participated in the Silkworm Genome Project, a joint research of China and Japan. The silkworm genome was the first sequence to be completed for the lepidopterans. By accelerating the production technology development of medical proteins and silk with new functions through analysis, it may contribute to developments of new agrichemicals and insect industry.

We collaborated with the groups from the University of Tokyo and the Kyoto Institute of Technology, and identified the seven transmembrane helix receptors from the silkworm genome and clarified the family distribution. Particularly, we found several characteristics unique to silkworms compared to other insects (drosophila, anopheles, and honeybee) concerning olfactory and gustatory receptors[12].

For this project also, it was necessary to modify the SEVENS pipeline for insects. We conducted studies of sequence resemblance score when pasting the known gene onto the genome, survey of additional extension for upstream and downstream, and the hidden Malkov modeling for common sequences seen only in the insect olfactory receptor. Also, we introduced a new protocol since the aim was to maximize the number of identified genes. In ordinary pipelines, when the known genes are used as seeds, a greater number of gene candidates emerge including new genes. Therefore, by using theses new genes as initial seeds of the pipeline, the number of new genes will increase. This is repeated sequentially until the number of predicted genes settles out (recursive computation). We identified 66 olfactory receptors. Among these, we identified 18 expressions of new receptors, and the odorous material (cis-Jasmone) that attracts the silkworm to mulberry leaves and its receptor were identified for the first time in the world. This became a world-class result in the field of biology, and was published in *Current Biology*[13].

The pipeline for insects and recursive computation protocol are reflected in the current SEVENS.

4.6 Current results, SEVENS and GRIFFIN

As of 2009, SEVENS stores 24,545 genes for 43 eukaryotes, under the support of Grant-in-Aid for Scientific Research (Grant-in-Aid for Publication of Scientific Research Results). It is an integrated DB where various kinds of functional and structural information are visually presented and organized in hierarchical manner. The technologies improved in the joint researches are fed back, and the current information volume is abundant. Figure 4 shows the web page (http://sevens.cbrc.jp) of the current SEVENS.

The top page shows the list of the eukaryotes, and when an organism type is selected, the search page is shown. One can jump to the GPCR detailed analysis page from the chromosome map, the phylogenetic icon, or the search condition entry form. From the detailed analysis page, such information as the coordinates of selected GPCR, exon sequence, sequence resemblance search, gene expression pattern, ligand binding, G protein binding, composition of amino acid sequence, predicted transmembrane helix region, functional motif region, domain region, region predicted to be indeterminate structure (disorder region), exon-intron boundary, pseudo-gene, new gene, and 3D structure modeling can be viewed.

GRIFFIN that was developed for functional prediction can be used on the web (http://griffin.cbrc.jp/). When the molecular weight of the ligand and GPCR sequence are entered, the bonding G protein is predicted. The ligand molecular weight can be set in arbitrary steps for a certain value. The step-by-step ligand molecular weight setting is useful for the prediction of bonding G protein with an orphan receptor whose ligand is unknown.

5 Jump again: future research development

5.1 Understanding high-order biological phenomena

Up to this point, we placed weight on the functional analysis of individual genes from a comprehensive perspective, but in the future, research to understand high-order biological phenomena based on the entire gene network is necessary.

From this perspective, we started working on the system that involves the olfactory receptors that dominate the majority of the mammalian GPCR. The olfactory system induces memories and emotions through multitudinous combination of odor molecule types. Therefore, if this system can be understood systematically, it may lead to research for producing an environment that makes people feel pleasure by blending certain odor molecules.

The electric activation signals from all of the several hundred olfactory receptors that respond to diverse odor molecules are integrated to form 2D patterns (odor map) on the olfactory epithelium. We would like to understand the spatiotemporal cause-and-effect relationships among odor molecules, receptors, cells, and the odor map. Specifically, we are planning to develop a program to predict the activation of all olfactory receptors against odor molecules (activation array), and apply this to all olfactory receptors of humans and mice.

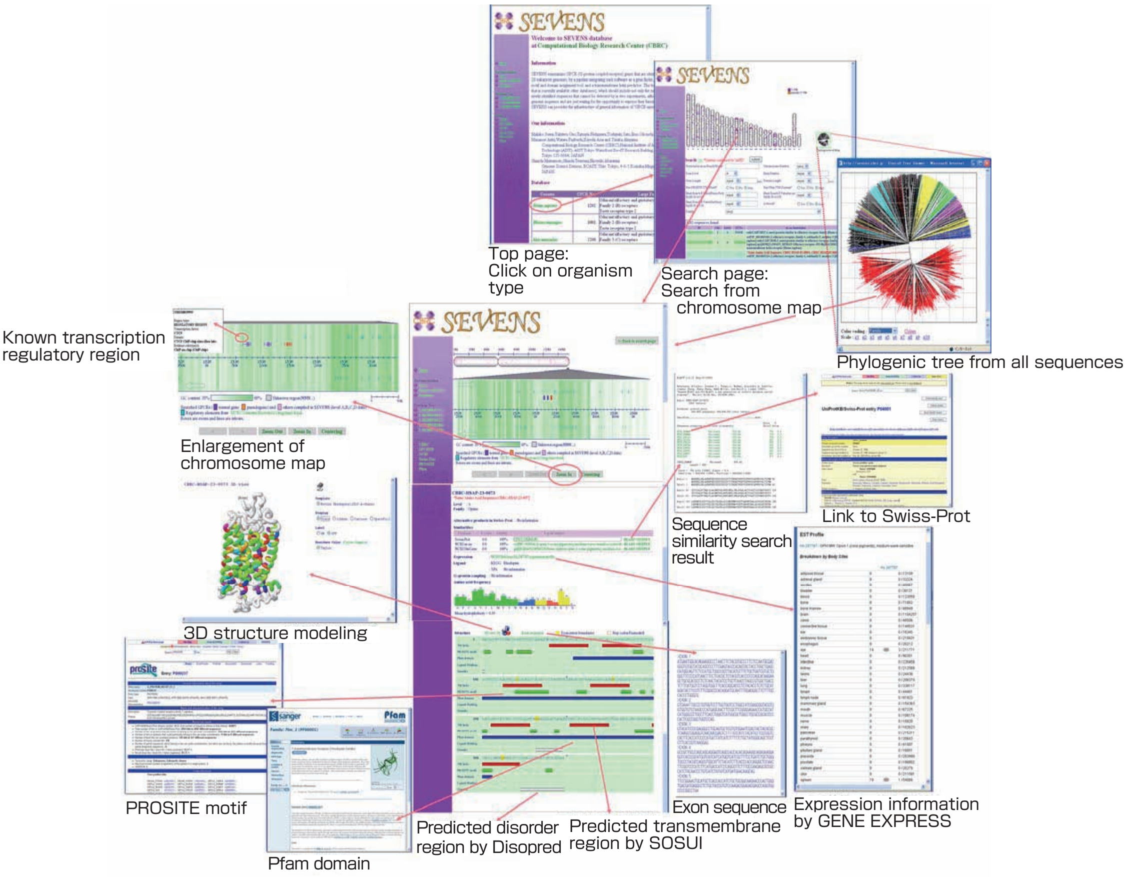
SEVENS
Top page:
Click on organism type
Search page:
Search from chromosome map
Phylogenic tree from all sequences
Known transcription regulatory region
Enlargement of chromosome map
Sequence similarity search result
Link to Swiss-Prot
3D structure modeling
PROSITE motif
Pfam domain
Predicted disorder region by Disopred
Predicted transmembrane region by SOSUI
Exon sequence
Expression information by GENE EXPRESS

Fig. 4 Current SEVENS database (http://sevens.cbrc.jp).

We already possess all olfactory receptors in SEVENS. We believe we can conduct a response simulation of all olfactory receptors against odor molecules by modifying GRIFFIN.

5.2 New research phase of GPCR

It is necessary to consider the recent rapid advancement in the research on the 3D structure of GPCR. For a long time, the only protein to have its 3D structure deciphered was bovine rhodopsin, and this was used as a mold of the model structure to be analyzed, as a matter of fact, in drug discovery research. However, in 2007 to 2008, the GPCR structures of different families[14][15] were determined one after another, and the conventional research method is expected to change rapidly.

From the new 3D structure, it was found that the differences in the structures of the ligand binding site and the G protein binding site were spread out and could not be ignored between families. It was indicated that it is necessary to determine the 3D structures of all major GPCR families as molds. However, such immediate structure determination is quite difficult as both expression and crystallization have become bottlenecks. Therefore, structural information must be obtained through a different approach from the 3D structure determination. It is important to extract and overview the information that reflects the 3D structure for each family at sequence level, and SEVENS can be used for this very purpose.

5.3 Development of an integrated database

Although the databases containing bioinformation are the foundation of the life science research, their utility is low since they are scattered around in various research institutions. Therefore, the government is working on a system for integrating and managing the various DBs (for example, the Integrated DB Project of MECSST and Ministry of Economy, Trade and Industry). In the future, SEVENS must be designed for such integrations. It is necessary to take measures to completely automate updates for permanent maintenance and management, and yet maintain high data reliability.

6 Discussion

6.1 Research scenario: Cyclically developing Project structure

The results accomplished and the future developments of the Project were presented in the above chapters. It was stated earlier that it is difficult to write a “correct” research scenario into the far future since the advancement of research in the life science field is quick. Yet in retrospect, I think the research developed extremely efficiently. The Project started in 2000, and the first phase of *Full Research* was from the development and publication of the comprehensive DB of GPCR. However, this phase was the “hop” (*Type 1 Basic Research*) of the larger research development phase, and this was followed by the cyclic development of joint research in the manner of *Type 2 Basic Research*, joint research in the manner of *Product Realization Research*, and continues to the present (Fig. 5(a)).

Why did it take such a development form rather than being linear? One could think of the following reasons. First, as described in chapter 2, the time needed for results is short in the bioinformatics fields, and each research phase of Fig. 5(a) tends to be small projects that are resolved in 1~2 years. Considering such small-scale research directions as small vectors, the vector that is synthesized from those small vectors and the direction of life science in its entirety determines the direction of the whole project. The determination of the direction is accomplished at each phase. Next, since the direction of the life science field moves in cyclically developing manner together with dramatic technological development, it will continue to develop under such influence.

What was the driving force that propelled these small vectors without interruption to the present? Following is a list of the factors, and I think these worked as shown in Fig. 5(b) to determine the direction of research.

1) Core technology matured over a long-term

Over 8 years were spent on the Project. Normally, a project spans over about 5 years, and orders to stop the research could have been given at any time. However, in our case, the stage of research continued to rise through the long-term maturation of the core technology. The essential factor that prevented the interruption of the cyclically developing structure is the fact that SEVENS itself won trust by diligently building up improvements in gene identification pipeline, DB, and program. While there are many DBs that are completed, written up in a paper, and are never maintained afterwards, the fact that SEVENS is updated in accordance to the demand of the moment has become a brand power, and I think this is the reason we are getting offers for joint researches.

2) Close collaboration with experimental researchers

The bioinformatics technology is able to process large amounts of data in a short period and produce results. However, whether the results are meaningful or not must be demonstrated in experimental research. By obtaining feedback of the demonstrated results, the parameters set in elemental technology can be modified to move in a better direction. On the other hand, the experimental researchers can use the predictions to modify their experimental system to one with lower risk and cost. In our Project, we discussed extensively with experimental researchers through various joint researches. Feedback in both directions occurred several times, and the improvements of analysis and prediction technologies were accelerated. Our research center does not conduct experiments, but we feel it is necessary to collaborate with experimental researchers at all times in all researches in the future.

3) Place of incubation

The Project started in 2000, around the time of the establishment of the Computational Biology Research Center. It was not necessarily a good start. It was totally new without any model, and we tried to figure out how to start things and groped along the way. Of course, we did have some idea of how the Project should progress. While I had an image of “this can be done by doing this” as a researcher who has been studying the cell membrane protein for a long time, I could have never come up with a method toward specific realization on my own. Working with Dr. Akiyama, a specialist in parallel computing environment, and Dr. Asai, an expert in mathematical models, a powerful analysis could be done using a parallel supercomputer environment and advanced mathematical methods. To this day, discussions with other researchers are inspirational in various scenes. This could not have been possible if it were not done at the Computational Biology Research Center where researchers of various backgrounds are gathered in one place, and I am grateful for this opportunity.

6.2 Achievement of the research objectives

The objective of the Project at its start was to present information that may contribute to the design of experiments by predicting the experimental result to reduce major experimental risks using the bioinformatics technology in GPCR research. Compared to 2000, proteins other than GPCR such as Kinase and protein complex formation inhibitors dominate the higher percentage as drug discovery targets. However, the importance of GPCR has not faded, and the number of academic papers on GPCR is increasing with the increase of bioinformation. Did we achieve our objective in all this?

SEVENS has already analyzed genes that show potential expression in the body as well as GPCRs whose expressions have been confirmed by experiments. Therefore, it is unique since it is capable of a truly comprehensive analysis. We are certain that it can contribute greatly to the general understanding of GPCR and the related drug discovery. Whether it does contribute or not can be indicated by how often the developed tools were used and how much feedback was received. Currently, it is linked to the portal sites of international journals and Integrated DB Organization of MEXT and METI. There are on average about 1,000 serious accesses per month from companies and government organizations in Japan and other countries (such as United States, Germany, France, Brazil, Spain, Italy, and Taiwan). It is also reviewed in international literatures[16][17] as one of the major web DB for drug discovery. GRIFFIN is competing for the top position as a web tool for predicting the G protein

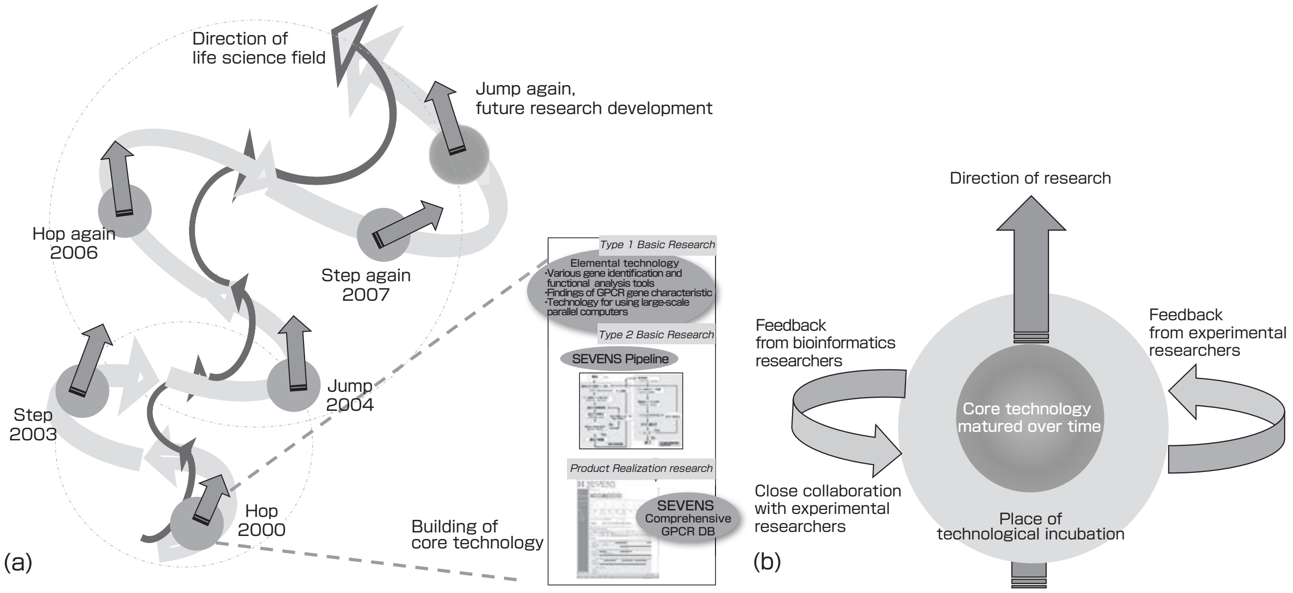

Fig. 5 Conceptual diagram of the research project structure that undergoes cyclical development.

(a) The process from the development to publication of the comprehensive DB for GPCR at the beginning of the Project was *Full Research* in a small scale. This phase was incorporated into the *Type 1 Basic Research*-like phase (hop, hop again) of the greater research development phases, followed by the cyclical development of *Type 2 Basic Research*-like joint research (step, step again) and *Product Realization Research*-like joint research (jump, jump again). This is a form that continues to develop through the interaction of directions of joint researches at each step and the direction of the life science fields that continue to advance rapidly.

(b) Relationship of the three factors that serve as the driving force of the joint research at each step. (1) The core technology that matured over a long period continues to grow and mature further in (2) the research environment that nurtured technological incubation. The cyclic movement (3)with feedback from the close collaboration between bioinformatics researchers and experimental researchers, based on (1) and (2), determines the direction of the vector of joint research. This is much like the determination of the axis direction by the rotation of a gyro.

bond, and this is also reviewed in international literature[18].

As described in chapter 4, various joint researches by collaboration among industry, academia, and government were conducted under cyclic development, and yielded important results. Although this was inconceivable initially, I am surprised that the research developed extremely efficiently in retrospect. At the beginning of the Project, there were mainly joint researches with companies, but joint researches with academia increased in the past 3 years. This shows that the users of SEVENS are increasing and covering wider areas. It is a joy to hear from many experimental researchers of pharmaceutical companies and universities that I meet for the first time at scientific conferences that they use SEVENS or GRIFFIN and that it is very useful in analyzing new genes. Looking back, the initial objectives were achieved to some degree, and I shall give a self-evaluation as being satisfactory.

The SEVENS project will continue to develop in the future. Based on the functional data accumulated over a long time, we wish to produce results that lead to the clarification of high-order biological phenomena in which GPCR is involved through advanced collaboration with experimental researchers.

Acknowledgements

This Project is a joint research with many people. Yutaka Akiyama (Tokyo Institute of Technology; former Director of CBRC), Kiyoshi Asai (The University of Tokyo; Director of CBRC), Masanori Arita (The University of Tokyo), Professor Hiroyuki Aburatani (The University of Tokyo), Tomoyuki Sato (Mizuho Information and Research Institute), and Ikuo Ohkochi (Mizuho Information and Research Institute) helped us incorporate the GPCR gene identification technology. Takatsugu Hirokawa (research team leader, CBRC) and Yukimitsu Yabuki (Information and Mathematical Science Laboratory, Inc.) helped us with the development of GRIFFIN. Wataru Fujibuchi (research team leader, CBRC), Tatsuya Nishizawa (Information and Mathematical Science Laboratory, Inc.), and the students of Nara Institute of Science and Technology helped us in the comparative genome analysis for GPCR. We are truly grateful to the above-mentioned collaborators.

References

[1] E. S. Lander *et al.*: International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome, *Nature*, 409, 860-921 (2001).

[2] A. Shenker: G protein-coupled receptor structure and function: The impact of disease-causing mutations, *Baillieres Clin. Endocrinol Metab.*, 9, 427-451 (1995).

[3] O. Gotoh: Homology-based gene structure prediction: Simplified matching algorithm using a translated codon, (tron) and improved accuracy by allowing for long gaps, *Bioinformatics*, 16, 190-202 (2000).

[4] http://genedecoder.cbrc.jp/

[5] http://hmmer.janelia.org/

[6] T. Hirokawa, S. Boon-Chieng and S. Mitaku: SOSUI, Classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, 14, 378-379 (1998).

[7] M. Suwa, T. Sato, I. Okouchi, T. Kumagai, M. Arita, K. Asai, Y. Akiyama, S. Matsumoto, S. Tsutsumi and H. Aburatani: SEVENS, *Nucleic Acids Research*, 31, Online summary paper (http:// www3.oup.co.uk/nar/ database/ summary 373), (2003).

[8] Y. Terashima, N. Onai, M. Enomoto, V. Poonpiriya, T. Hamada, K. Motomura, M. Suwa, T. Ezaki, T. Haga, S. Kanagasaki and K. Matsushima: Pivotal function for cytoplasmic protein FROUNT in CCR2-mediated monocyte chemotaxis, *Nature Immunology*, 6, 827-835 (2005).

[9] T. Muramatsu and M. Suwa: Statistical analysis and prediction of functional residues effective for GPCR-G-protein coupling selectivity, *PROTEIN Engineering Design & Selection*, 19, 277-283 (2006).

[10] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai and M. Suwa: GRIFFIN, a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model, *Nucleic Acid Research*, 33, W148-W153 (2005).

[11] Y. Ono, W. Fujibuchi and M. Suwa: Automatic gene collection system for genome-scale overview of G-protein coupled receptors in eukaryotes, *Gene*, 364, 63-73 (2005).

[12] Q. Xia *et al.*: Silkworm genome consortium, The genome of a lepidopteran model insect, the silkworm Bombyx mori, *Insect Biochemistry and Molecular Biology*, 38, 1036-1045 (2008).

[13] K. Tanaka, Y. Uda, Y. Ono, T. Nakagawa, M. Suwa, R. Yamaoka and K. Touhara: Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile, *Curr. Biol.*, 19, 881- 890 (2009).

[14] M. A. Hanson and R. C. Stevens: Discovery of new GPCR biology, one receptor structure at a time, *Structure*, 17, 8-14 (2009).

[15] D. T. Lodowski, T. E. Angel and K. Palczewski: Comparative analysis of GPCR crystal structures, *Photochem Photobiol.*, 85425-85430 (2009).

[16] L. X. Yao, Z. C. Wu, Z. L. Ji, Y. Z. Chen and X. Chen: Internet resources related to drug action and human response: A review, *Applied Bioinformatics*, 5, 131-139 (2006).

[17] L. J. Zhi, L. Z. Sun, X. Chen, C. J. Zheng, L. X. Yao, L. Y. Han, Z. W. Cao, J. F. Wang, W. K. Yeo, C. Z. Cai and Y. Z. Chen: Internet resources for proteins associated with drug therapeutic effects, adverse reactions and ADME, *Drug Discovery Today*, 8, 526-529 (2003).

[18] A. Daskalaki ed.: Handbook of Research on Systems Biology Applications in Medicine, Vol I (Medical Information Science Reference Press) (2009).

Authors

Makiko Suwa
Principle Research Scientist of Computational Biology Research Center, AIST. Completed studies at the Graduate School of Science and Engineering, Aoyama Gakuin University in 1986. Doctor (Science). Worked as technical official, associate researcher of education, Faculty of Technology, Tokyo University of Agriculture and Technology; senior investigator, Helix Research Institute Inc.: chief researcher, Electrotechnical Laboratory, Agency of Industrial Science and Technology; and research team leader and deputy director, Computational Biology Research Center, AIST. Assumed the current position in 2007. Specialties are bioinformatics and biophysics. In charge of the overall coordination for the Project described in this paper.

Yukiteru Ono
Manager of 8th Business Group, Biolife Science System Enterprise Division, Information and Mathematical Science Laboratory, Inc. Completed the master's course in biology at the Graduate School of Science, Nagoya University in 1994. Worked at JASTEC Co., Ltd. and assumed current position in 2001. Specialty is bioinformatics. For this paper, worked mainly on the web interface development for SEVENS and on the implementation of improvements resulting from the joint researches.

Discussions with Reviewers

1 Emphasis points on how the research was carried out

Comment (Motoyuki Akamatsu, Human Technology Research Institute, AIST)

As a *Synthesiology* paper, it is expected that the content will be about "bioinformatics strategy" as mentioned in the title. "Strategic" means that the research was carried out by setting a goal and laying out the research scenario (process) beforehand. If the authors intentionally set up ways to conduct cyclical research, please describe them. If the authors did not intend to do so but things developed spontaneously, I think you should describe what were the conditions necessary for such cyclical development of the DB to occur. Also, I think the main point of this paper is the description of the cyclical development process of the DB, so I think the point will be easier to be understood if you include a diagram of this development process.

Comment (Hideyuki Nakashima, Future University Hakodate)

I think it better to stress the points on your research method for the sake of general readers (researchers of other fields).

Answer (Makiko Suwa)

The bioinformatics strategy described in the text was not necessarily conducted by setting up a research scenario beforehand and then following the road map. Rather, in retrospect, I feel the research developed extremely efficiently regardless of my intent, and therefore I focused on the driving forces that are unique to bioinformatics and that enabled such development.

I think the overall flow of the development of the research project has a dynamic form where there is an upward spiral of the interactions of the direction of individual research that progresses with multiple factors (core technology that matured over time, close collaboration with experimental researcher, environment that nurtures technological incubation, etc.) as driving forces, and the direction of the life science field that evolves extremely quickly. (This development process is shown in Fig. 5(a) and (b)).

I think this form is the result of the characteristics of bioinformatics: it can set diverse directions depending on the situation since it is not strongly limited by research targets; and the individual researches are resolved in 1~2 years since the period required from basic research to application and realization is short.

2 Title

Comment (Motoyuki Akamatsu)

Please consider a title that represents the content from the synthesiological perspective. It must show that this is a paper that discusses research progress where the DB moved in an upward spiral through joint research.

Answer (Makiko Suwa)

The first title "Search and functional analysis of drug discovery target GPCR - Bioinformatics strategy" represented the content of the research, but as you indicated, it did not explain how the bioinformatics strategy is related to the entire Project when seen from a synthesiological perspective. Therefore, to clarify that point, I changed the title to "A bioinformatics strategy to produce a cyclically developing project structure - Comprehensive functional analysis of the drug design target genes."

3 *Type 2 Basic Research*

Comment (Motoyuki Akamatsu)

At the end of paragraph 1 of section 3.2 "Gene identification and functional analysis pipeline," it says that these combined researches can truly be called *Type 2 Basic Research*. If possible, can you explain what points you consider to be *Type 2 Basic Research*?

Answer (Makiko Suwa)

The description you indicate is the part about the development of gene identification and functional analysis pipeline. This can be called *Type 2 Basic Research* because the work is done from the perspective where the elemental programs that are established after accumulation of basic research are combined and systematized, and these are controlled and applied to the subject. I added this explanation.

4 Bioinformatics

Comment (Hideyuki Nakashima)

The explanation of "bioinformatics" on the first page emphasizes the aspect of information technology as a tool for biology. Certainly, that aspect is strong in this paper, but CBRC has emphasized that IT is not merely a tool I think it is better to add the points that the way of thinking and the approach of the information science are also important.

Answer (Makiko Suwa)

I agree that the description you indicate gives the impression that bioinformatics is "merely a tool". It is because we intend to emphasize the advantages of bioinformatics technology, when looking at it from the perspective of reducing the difficulty in experimental research approach. Therefore, I modified the description so that it indicates first the general definition of bioinformatics and after, the above advantage as one part of the whole picture.

Bioinformatics is a wide-ranging discipline with a collection of researchers of varying backgrounds. I feel that the understanding of the definition and the aspect one handles are diverse as they depend widely on the background of the researcher. In my case, I place emphasis on biological findings because I have a biophysics background.

Therefore the trial-and-error process tends to be unrefined work, depending heavily on the research subject. The ideations for "using a tool" such as in which order and how to combine the programs for a particular subject are essential, and that aspect became apparent in the text.

Our approach is different from the information science approach where a beautiful system is applied to whatever subject, but our unrefined approach is also accepted. I believe such diversity gives breadth of expanse in the field of bioinformatics.