An Automated System for Finding Seven Transmembrane Helix Receptors from Human Genome

Makiko Suwa1
m-suwa@aist.go.jp
Toshiyuki Sato1,2
sato-toshiyuki@aist.go.jp
Ikuo Okouchi1,2
i-okouchi@aist.go.jp
Masanori Arita1
m-arita@aist.go.jp
Shinichi Matsumoto3
Matsumoto@genome.rcast.u-tokyo.ac.jp
Shuichi Tsutsumi3
shuichi@genome.rcast.u-tokyo.ac.jp
Hiroyuki Aburatani3
haburata-ty@umin.ac.jp
Kiyoshi Asai1
asai-cbrc@aist.go.jp
Yutaka Akiyama1
akiyama-yutaka@aist.go.jp
1 Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan
2 Center for Computational Science and Engineering, Fuji Research Institute Corporation, 2-3 Kandanishiki-cho, Chiyoda-ku, Tokyo 101-8443, Japan
3 Genome Science Division, Research Center for Advanced Science and Technology (RCAST), University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan

Keywords: aberrant splicing, database, point mutation, scanning model

1 Introduction

Seven-transmembrane-helix receptors (7-TMR), most of which are G-protein-coupled receptors are one important target of drug gene discovery. They are experimentally well characterized and are known to mediate the signal transduction induced by ligand binding. Despite the enormous amount of available data on cDNA, EST, and genome sequences, however, the overview of the large-scale collection of the 7-TMR family sequences is unclear. We therefore developed an automated system for discovering 7-TMR sequences in the whole human genome [5, 6, 7, 8]. It integrates tools for gene finding [2], sequence searching [1], motif and domain assignment [3, 4], transmembrane helix prediction, and gene quality refinement and can detect sequences of multiple exons or remote homologues that cannot be detected by conventional sequence search tools. By carefully assessing each component, we predicted the sequences of 7-TMR genes from datasets with four confidence levels.

2 Method and Results

The automated discovery system is composed of three stages.

(1) Gene prediction stage: We used human genome sequences from finished and draft sequences [5, 6, 7, 8]. We predicted amino acid sequences from genome sequences using two methods: the prediction of all possible combinations of 6 open reading frames between possible initial and stop codon signatures, and the gene finding program (GeneDecoder [2] with Hidden Markov Model) to translate the entire structure of multiple exon sequences.

(2) Sequence screening stage: We screened 7-TMR candidates by analyses of sequence search [1], the motif (PROSITE [3]), the domain (PFAM [4]) assignment, and transmembrane helix prediction. In previous of this analysis, we first evaluated each sequence analysis program by using a reference dataset with sequences from the SWISS-PROT database. We focused on two threshold settings: the best specificity and the best sensitivity. Then we prepared several levels of datasets, given from outputs...
by analysis with various kinds of combination of the best specificity and best sensitivity thresholds. 

(3) Gene quality improving stage: We adjusted the redundancy of predicted sequences when they are overlapped at the same genetic position (same chromosome number, contig number and relative position at the contig). While we merged separated fragment sequences when they can be linked by known sequence (as a intermediate sequence) on the same genetic position.

The final output was 7-TMR candidates datasets with four confidence levels, in which 888 candidates at the highest confidence level and 2,248 candidates at the lowest confidence level. Throughout the four levels of dataset, it was shown that chromosome 11 has the largest number of 7-TMR candidates, while chromosomes 21 and Y, on the other hand, have few candidates. Further analysis of the most accurate 888 candidates revealed that the largest family was odorant/olfactory and gustatory receptors. Throughout all chromosomes the percentages of genes for odorant/olfactory and gustatory receptors were high and especially in chromosome 11 more than 90% of the candidates were for this kind of genes. The other major categories with more than 20 members were adrenergic, dopaminergic, and serotonergic receptors, chemokines and chemotactic receptors, family 3(C) receptors, and family 2 (B) receptors.

An advantage of this system is that it is applicable to any species (mouse, drosophilae, microviral germs, etc.) because it is tuned by known 7-TMR sequences, including those of all species in the SWISS-PROT database. Thus the next step in our research will be the comparative genomic study of 7-TMR genes.

References


