Published in final edited form as:

J Bioinform Comput Biol. 2014 February ; 12(1): 1350019. doi:10.1142/S0219720013500194.

FUNCTIONAL UNDERSTANDING OF THE DIVERSE EXON-INTRON STRUCTURES OF HUMAN GPCR GENES

DOROTHY A. HAMMOND,

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

VICTOR OLMAN, and

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

YING XU

Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, and College of Computer Science and Technology, Jilin University, China

DOROTHY A. HAMMOND: darabah@uga.edu; VICTOR OLMAN: ludovicol@gmail.com; YING XU: xyn@bmb.uga.edu

Abstract

The GPCR genes have a variety of exon-intron structures even though their proteins are all structurally homologous. We have examined all human GPCR genes with at least two functional protein isoforms, totaling 199, aiming to gain an understanding of what may have contributed to the large diversity of the exon-intron structures of the GPCR genes. The 199 genes have a total of 808 known protein splicing isoforms with experimentally verified functions. Our analysis reveals that 1,301 (80.6%) adjacent exon-exon pairs out of the total of 1,613 in the 199 genes have either exactly one exon skipped or the intron in-between retained in at least one of the 808 protein splicing isoforms. This observation has a statistical significance p -value of 2.051762×10^{-9} , assuming that the observed splicing isoforms are independent of the exon-intron structures. Our interpretation of this observation is that the exon boundaries of the GPCR genes are not randomly determined; instead they may be selected to facilitate specific alternative splicing for functional purposes.

Keywords

gene structure; alternative splicing; G protein-coupled receptor; GPCR; exon-intron structures

1. Introduction

The G protein coupled receptors (GPCRs) are a large superfamily of transmembrane proteins with substantial applications to human therapeutics. At least 60% of the current drugs are targeted at the GPCR proteins.¹ They comprise at least five distinct families: glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin families; and the rhodopsin-like

Correspondence to: YING XU, xyn@bmb.uga.edu.

Supplementary Material

Supplementary Table 1:

<http://csbl.bmb.uga.edu/publications/materials/dhammond/S2.html>

Supplementary Table 2:

<http://csbl.bmb.uga.edu/publications/materials/dhammond/S1a.html>

GPCRs represent the largest family among the five. The current understanding is that rhodopsin-like GPCRs represent probably the most ancient GPCRs and the other GPCR genes are the evolutionary descendants of this family through gene and chromosomal duplications as well as other evolutionary events.² These transmembrane proteins can each sense distinct extracellular signals and then activate the relevant signaling pathways to trigger the execution of the designed cellular responses, where these signals include hormones, neurotransmitters, ions, photons and other stimuli. Different from many other gene (super) families, the GPCR genes have a large number of very distinct exon-intron structures even though their proteins are all structurally homologous. A simple examination of their gene structures reveals that the number of exons across different GPCR genes ranges from 2 to 42 with the exon length ranging from as low as 15 bps to well over 10K bps; and the intron length ranges from 4 bps to well over 80K bps.

The question we address in this short report is: *what may have contributed to the large diversity of the exon-intron structures among the GPCR genes?* To the best of our knowledge, no explanation has been offered in the published literature to this question. We have examined this issue from the perspective of alternative splicing.

Alternative splicing is a cellular process by which multiple mRNA transcripts can be produced from the same pre-mRNA through exon skipping, intronic region retention and alternative 5' or 3' ends, which is estimated to occur in approximately 92% of the human genes.^{3, 12, 13} As of now at least 50% of the GPCR genes have been experimentally validated to have functional protein splicing isoforms.⁴⁻⁶ Proteomic analysis further indicate that the majority of the functional protein species in human cells are in splicing isoforms.⁷ Substantial amount of information has been derived on the various splicing isoforms encoded in the human genome based on the RNA-seq data^{3, 8, 10, 11, 14-18}.

We have carried out a statistical analysis of all the 808 GPCR splicing isoforms with experimentally validated functional proteins of all the known GPCR genes with at least two validated splicing isoforms, totaling 199.⁹ The idea of our analysis is to show statistically that there is a strong correlation between the specific way in splitting the 199 GPCR coding messages into the current sets of exon-exon boundaries and the 808 known splicing isoforms. While this correlation does not necessarily imply causality relationships, we speculate that the constraints cast by the splicing isoforms needed by the diverse GPCR functions may have contributed to the splitting of the GPCR coding messages into the current exon-exon boundaries.

2. Results

The 199 GPCR genes under consideration have diverse ways of splitting their protein-encoding regions into a large set of very distinct exons and introns, as illustrated by one GPCR gene GPCR5C and its known splicing isoforms in Figure 1. Overall, the 199 genes consist of 1,613 adjacent exon-exon boundaries (or pairs). Out of these pairs, 1,203 each have exactly one exon skipped and 98 have the in-between introns retained in at least one of the 808 splicing isoforms. Overall, 1,301 out of the 1,613 (80.7%) exon-exon boundaries have been used in alternative splicing; conversely, an exon-exon pair is considered as not used by any splicing isoform if all the splicing isoforms each include both exons without their intron retained or none of the two exons. The statistical significance of this observation, assuming that the exon-skipping (skipping the whole exon) and intron-retention events are random, is 2.051762×10^{-9} , using the method given in the Methods section. This high statistical significance indicates that there is a strong correlation between all the 808 known splicing isoforms and the exon boundaries in the 199 GPCRs.

This analysis result suggests that individual exons and introns that are skipped or retained in splicing isoforms may have their own functional elements. Hence we have examined the possible functional roles of the 1,301 splicing events by checking if any of them have known functional motifs and by comparing their encoded-peptide locations in the secondary structures of their whole proteins. We found that (1) 100% of the skipped exons and the retained introns each cover an entire loop region, 17% of which also cover entire helices; and (2) the majority (75%) of the skipped exons and retained introns, for example, in the dopamine sub-family of the GPCR genes, contain known functional motifs, which have more annotated functional motifs than the other GPCR families/sub-families. We did not carry out motif searches for the whole set of the 199 GPCR genes but focused on this one sub-family as the detailed motif work will be reported in a separate paper (working manuscript under preparation). All the detailed data for these statistics are given in Supplementary Table 1.

The significance of the observed relationship between the exon boundaries and the known splicing isoforms is that for the first time, someone provided statistical evidence that the functional needs for the specific splicing isoforms may have cast the evolutionary constraints on the selection of the current exon boundaries. Clearly this warrants further investigation to address the highly important evolutionary question: what may have contributed to the selection of specific exon boundaries in a eukaryotic genome?

3. Material and Methods

A total of 205 human GPCR genes with multiple transcripts were retrieved from the HUGO Gene Nomenclature Committee (HGNC) database.¹⁹ Based on the information in Ensembl (ensembl 71 (e!71) - April 2013),⁹ 199 of the 205 genes each have at least two protein products and are not candidates for nonsense mediated decay, totaling 808 experimentally verified functional splicing isoforms. The 199 genes cover four out of the five major families of the GPCR genes: glutamate, rhodopsin, taste/fizzle and secretin (but not adhesion). The detailed list of these splicing isoforms along with the relevant functional motifs is available in Supplementary Table 2. The secondary structures of each transcript were predicted using the TMHMM program.²⁰ Motifs were determined using the expectation maximization technique as used by the MEME²¹ motif finding tool and a search of PROSITE.²²

On a per gene basis, we have assessed exon-exon pairs by counting the number of exon boundaries, totaling 1,613 adjacent exon-exon pairs. For each of these 199 genes the number of exon pairs with at least one exon removed or the intron between them retained in at least one of the isoforms yielded a total of 1,301. An exon-exon pair is considered as not used by any splicing isoform if all the splicing isoforms each include both exons without their intron retained or none of the two exons.

We have used the following procedure to assess the statistical significance in having observed that 1,301 distinct exon-exon pairs out of the total of 1,613 each have exactly one exon skipped or the intron in-between retained in the 808 splicing isoforms. Intuitively, this observation seems to be highly unexpected if exon boundaries are largely independent of alternative splicing. We consider each of the 1,301 exon-exon boundaries as having splicing signals. Our null hypothesis is that exon boundaries with splicing signals are independent of each other. For each splicing isoform, let S be the number of exon clusters present in the isoform, with a *cluster* being the maximum run of the consecutive exons without skipped exons or retained introns in the middle. Based on the null hypothesis all R-exon isoforms of the E-exon genes should have the equal probability to be formed. Then the density of S can

be calculated using the following formula based on the Theory of Series, which is a well-established model for studying series of runs since 1940's²³.

$$P(S|E, R) = \binom{R-1}{S-1} \binom{E-R+1}{S} / \binom{E}{R}$$

with E being the number of exons in the underlying gene and R the number of exons present in the current isoform. The *p*-value for each observed isoform can be calculated using

$$\sum_{S \leq S_{data}} P(S|E_{data}, R_{data})$$

with S_{data} , E_{data} and R_{data} being the S, E and R values for the current isoform, respectively. We then calculate the total *p*-value as the geometric average of the *p*-values across all the 808 isoforms, giving rise to a *p*-value = $2.051762 * e^{-09}$ for the whole set of splicing isoforms. This implies that our null hypothesis is false, hence indicating that there is a strong relationship between the current exon boundaries in the 199 GPCR genes and the specific set of the 808 observed splicing isoforms. Note that this specific set of 808 splicing isoforms already largely defines the exon boundaries as the above formulation requires the consistency between the exon boundaries defined in the genes and the segments of the messages included in the splicing isoforms when we consider only two types of splicing events, whole exon skipping and intronic region retention.

4. Concluding Remark

The key finding of this study is that 80.7% of the adjacent exon-exon boundaries in the GPCR genes are strongly correlated with the known splicing isoforms as the statistical significance of this observation is very high at $2.051762 * e^{-09}$. The observed statistical relationship hints that the functional needs by specific splicing isoforms of the GPCR genes may have cast evolutionary constraints in the selection of the specific boundaries of the exons. Clearly this warrants further and detailed phylogenetic analyses to confirm the roles of splicing isoforms in the determination of the exon boundaries, which represents a highly important evolutionary and fundamental biological question. While the data for GPCR splicing isoforms are clearly incomplete, we believe that the dataset we used are large enough, and the statistical significance is high enough to warrant further study along this line of research in elucidation of the fundamental causes in the selection of the specific exon boundaries.

Acknowledgments

DAH is supported by National Institute of General Medical Sciences, National Institute of Health (grant number F31 GM089093) Ruth L. Kirschstein Predoctoral Fellowship. Our appreciation goes to these current and former colleagues of the Computational Systems Biology Laboratory, University of Georgia Athens for their help and discussions, specifically to Dr. Xizeng Mao, Dr. Fengfeng Zhou and Dr. Yanbin Yin.

Funding

This work was supported by National Institute of General Medical Sciences, National Institute of Health [grant number GM089093] Ruth L. Kirschstein Predoctoral Fellowship.

References

1. Kriechbaumer V, Nabok A, Widdowson R, Smith DP, Abell BM. Quantification of ligand binding to G-protein coupled receptors on cell membranes by ellipsometry. *PLoS One*. 2012; 7(9):e46221. [PubMed: 23049983]
2. Rompler H, Stauber C, Thor D, Schulz A, Hofreiter M, Schoneberg T. G protein-coupled time travel: evolutionary aspects of GPCR research. *Molecular interventions*. 2007; 7(1):17–25. [PubMed: 17339603]
3. Mironov AA, Fickett JW, Gelfand MS. Frequent alternative splicing of human genes. *Genome Res*. 1999; 9(12):1288–93. [PubMed: 10613851]
4. Nakao M, Barrero RA, Mukai Y, Motono C, Suwa M, Nakai K. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res*. 2005; 33(8):2355–63. [PubMed: 15860772]
5. Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet*. 2002; 30(1):13–9. [PubMed: 11753382]
6. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*. 2003; 302(5653):2141–4. [PubMed: 14684825]
7. Collins FS, Lander ES, Rogers J, Waterston RH, Conso IHGS. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431(7011):931–45. [PubMed: 15496913]
8. Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? *Trends Genet*. 2004; 20(2):68–71. [PubMed: 14746986]
9. Strozzi F, Aerts J. A Ruby API to query the Ensembl database for genomic features. *Bioinformatics*. 2011; 27(7):1013–4. [PubMed: 21278190]
10. Barta A, Schumperli D. Editorial on alternative splicing and disease. *RNA Biol*. 2010; 7(4):388–9. [PubMed: 21140604]
11. Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell*. 2009; 136(4):777–93. [PubMed: 19239895]
12. Hallegger M, Llorian M, Smith CW. Alternative splicing: global insights. *Febs J*. 2010; 277(4):856–66. [PubMed: 20082635]
13. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. *Gene*. 2013; 514(1):1–30. [PubMed: 22909801]
14. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochim Biophys Acta*. 2009; 1792(1):14–26. [PubMed: 18992329]
15. Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, et al. Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. *Mol Cell Biol*. 2011; 31(1):138–50. [PubMed: 21041478]
16. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Gene Dev*. 2010; 24(21):2343–64. [PubMed: 21041405]
17. Poulos MG, Batra R, Charizanis K, Swanson MS. Developments in RNA splicing and disease. *Cold Spring Harb Perspect Biol*. 2011; 3(1):a000778. [PubMed: 21084389]
18. Ranum LP, Cooper TA. RNA-mediated neuromuscular disorders. *Annu Rev Neurosci*. 2006; 29:259–77. [PubMed: 16776586]
19. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res*. 2013; 41(Database issue):D545–52. [PubMed: 23161694]
20. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol*. 2001; 305(3):567–80. [PubMed: 11152613]
21. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 1994; 2:28–36. [PubMed: 7584402]
22. Sigrist CJ, de Castro E, Cerutti L, Cucho BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. *Nucleic Acids Res*. 2013; 41(Database issue):D344–7. [PubMed: 23161676]

23. Mood AM. The Distribution Theory of Runs. *Ann Math Statist.* 1940; 11(4):367–392.

Biographies

Dorothy A. Hammond is a Ruth L. Kirschstein Predoctoral Fellow and a PhD candidate in Bioinformatics at the University of Georgia, Athens. Her research interest is in G Protein-Coupled Receptors and alternative splicing. After receiving her Master of Public Health in Epidemiology from the Texas A & M University's School of Rural Public Health, she joined the Computational Systems Biology Laboratory at the University of Georgia in pursuit of a doctorate degree under the supervision of Prof. Ying Xu.

Victor Olman holds a PhD in Statistics from the Institute of Cybernetics of Estonian Academy of Sciences, Tallinn. A senior research scientist at the University of Georgia, his research interest is in cluster analysis, gene expression data analysis and biological pathway construction, and mathematical statistics.

Ying Xu is a Regents and Georgia Research Alliance Eminent Scholar and Professor in the Biochemistry and Molecular Biology Department at the University of Georgia (UGA). He received his Ph.D. degree in theoretical computer science from the University of Colorado at Boulder in 1991. Between 1991 and 1993, he was a visiting assistant professor at Colorado School of Mines. He started his bioinformatics career in 1993 when he joined Ed Uberbacher's group at ORNL to work on the GRAIL project, where he worked till 2003. During this period, he was a research associate, staff scientist and senior staff scientist and group leader. He joined the University of Georgia (UGA) in 2003 as an endowed chair professor and served as the founding director of the UGA Institute of Bioinformatics till 2011. His current research interests are mainly in cancer systems biology and microbial systems biology.

Figure 1. Shows a graphical representation of splicing variants of GPRC5C. Shaded bars indicate exons and dash lines indicate intronic regions. Top purple gene structure shows the entire gene region. Bottom turquoise graphs show the selected exons for the various mRNAs transcripts with resulting protein products for the gene. Each transcript is represented by its exon-intron structures exhibiting the inclusion and/or exclusion of certain exons. See legend for detailed color annotation.