

# Statement of Research Interests

J. Gregory Caporaso

gregcaporaso@gmail.com

www.caporaso.us

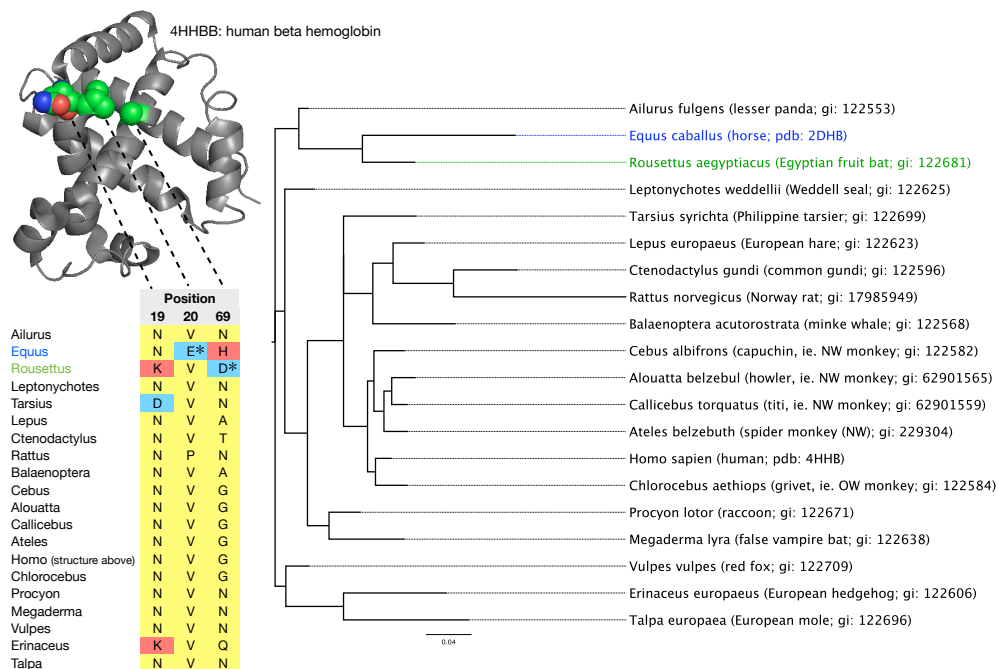
I am a computational biologist, primarily interested in studying and understanding the evolution of genes, proteins, genomes, and species. My recent research has focused on comparing orthologous protein sequences to make inferences about physical interactions within and between proteins, and specifically identifying pairs of positions which covary in multiple sequence alignments (MSA). In addition to specific research goals related to across-species sequence covariation, discussed in the following paragraphs, I would like to expand my studies to the effects of variation between members of the same species, including differences in genome sequence, genome content, microbiome and microbiota. Understanding diversity within and between species allows us to better understand present-day biological systems and thereby develop novel techniques to improve human health, avoid or respond to ecological threats, and increase our basic understanding of the processes and origin of life.

Patterns of covariation between positions in an MSA are thought to arise from phylogeny, stochastic variation, and more interestingly, coevolution of the involved positions. Identifying pairs of protein positions which coevolve is a sought-after goal as it is expected to imply physical or allosteric interactions between the positions. The challenge is to identify covariation arising from coevolution, while ignoring covariation arising from phylogeny (and to a lesser extent stochastic covariation, which is easier to control for). The next steps in this research will involve the continuation of several on-going projects, discussed below, in addition to new projects aimed at improving our understanding of the process of coevolution in protein sequences and our ability to identify truly coevolving positions in proteins.

I have recently shown that explicitly incorporating phylogeny into a coevolution metric may not be the best approach to control for phylogenetic covariation [1]. I hypothesize that this is because patterns arising from phylogeny and coevolution should be expected to look similar, as truly coevolving positions are likely to be under relatively strong selective pressure and therefore will infrequently incur substitutions. By compiling data from pairs of positions known to coevolve (e.g., base pairing positions in ncRNAs, stacked positions in protein alpha helices, or positions involved in compensatory mutation relationships) I would next like to test the hypothesis that patterns of covariation associated with coevolution mirror patterns of covariation arising from phylogeny (e.g., covarying positions in pairs of proteins thought not to interact). I suspect that this study will allow for an increased understanding of the process of coevolution between positions in proteins. Additionally, by elucidating any differences that do exist between these patterns, this study may suggest better algorithms for discriminating covariation patterns arising from phylogeny versus coevolution.

The assumption of independent evolution of positions in proteins is robust in many applica-

Figure 1: E20 and D69 (starred), associated with pathogenicity in humans, are wild-type in horse and fruit bat, respectively. If these otherwise-pathogenic deviations are alternately compensated by H69 and K19, pairwise coevolution algorithms would have difficulty detecting these relationships.



tions, and coevolution algorithms, which are designed to identify positions which do not evolve independently, frequently identify fewer position pairs than would be expected. Why do we not observe more pairs of protein positions evolving in a dependent manner when we know that selective pressures act on the phenotypic effects of mutations, which emerge from interactions between protein positions? I suspect one reason is that coevolution algorithms focus on pairwise interactions, when in many cases it appears that the positions involved in an interaction may change while conserving the interaction. For example, Figure 1 presents a putative example of this phenomena that I identified in beta hemoglobin. In the MSA, two substitutions (with respect to the human sequence) that are known to be pathogenic in human are present as wild-type in other species: V20E, present in horse; and G69D, present in fruit bat.

Kondrashov *et al.* [2] suggested that the G69H substitution present in horse was likely a compensatory mutation, which allowed the otherwise-deleterious E20 to be tolerated. Thus conservation of the interaction between positions 20 and 69 is more important than conservation of residues at those positions.

In reviewing data that I compiled to study this compensatory mutation relationship, I observed that the human-pathogenic D69 is present as wild-type in the Egyptian fruit bat sequence but that there is no compensatory change at position 20, its suspected interaction partner. I suspect that in this case an interaction with a residue at a nearby position may instead compensate for the V20/G69 interaction. One possible compensator is a relatively rare K19 residue present in the fruit bat sequence (although positions 19 and 69 may not be positioned to interact in the human structure, 4HHB). I suspect that the loss of an interaction between positions 20 and 69 in the human

sequence is associated with the human-pathogenic G69D mutation, and D69 may be compensated in fruit bat by a mutation at a different position, possibly K19. If this is the case, the interaction would still be conserved despite a change in the positions involved in the interaction. The non-pairwise nature of this compensatory mutation relationship would be problematic for most (if not all) current coevolution algorithms. I have observed other examples of this phenomena, and suspect that it is a common cause of the apparent independent evolution of protein sequence positions. In the next phase of my research on protein sequence coevolution, I would like to determine if this is indeed common by exploring approaches that I am developing to identify these types of situations algorithmically. This study would also result in an increased understanding of the process of coevolution between positions in proteins, and may lead to better algorithms for detecting protein coevolution.

## On-going research projects

### Protein coevolution

Through several interdisciplinary collaborations I am currently evaluating whether sequence cooccurrence and covariation can be applied to infer protein-protein interactions, identify tertiary structural interactions, and make informed decisions in protein engineering experiments.

### Sequence cooccurrence and covariation to infer protein-protein interactions

In collaboration with the Marcelo Sousa Biochemistry Laboratory at the University of Colorado at Boulder, I am attempting to identify intermolecular interactions based on sequence cooccurrence and covariation. I am evaluating three separate sets of putatively interacting bacterial proteins for sequence covariation, with the most mature study focusing on components of the Type VI Secretion System (T6SS).

The T6SS is a recently discovered protein complex used by pathogenic bacteria to transfer effector proteins from the pathogen's cytosol to the host's cytosol. Twenty-four *Salmonella enterica* serovar typhimurium proteins are potentially involved in the T6SS apparatus, but the specific protein-protein interactions important for complex formation are currently unknown. A brute-force approach to identify the interactions biochemically would involve testing 276 pairs of proteins for physical interactions. The two primary goals of this project are: (1) to identify the conserved pairwise protein-protein interactions in functional Type VI Secretion System complexes; and (2) to test the hypothesis that sequence covariation can suggest pairs of proteins that are likely to be involved in pairwise interactions, thereby reducing the cost (in dollars and person-hours) required to identify pairwise interactions in molecular complexes.

Of the twenty-four *Salmonella typhimurium* proteins, thirteen were found to frequently cooccur in bacterial gene clusters. My covariation analysis focused on these thirteen proteins, and predicted sixteen pairs of proteins (involving ten of the thirteen cooccurring proteins) to physically interact. Of these sixteen predictions, two have been biochemically confirmed; an additional five appear to be correct based on current information on the T6SS; eight involved predicted cytoplasmic proteins that have not been previously characterized; and only one prediction is inconsistent with the current information. This work has shed light on the organization of the T6SS complex, and all sixteen predictions are currently being evaluated biochemically. Upon completion of the biochemical aspects of this study by my collaborators in the Sousa lab, it will be possible to build a model of the T6SS, and to statistically test the hypothesis that sequence covariation can be used to infer

protein-protein interactions. This work was the subject of my poster presentation at the University of Colorado Denver Student Research Forum in January of 2009, where I won an Outstanding Research Award.

### **Sequence covariation to identify important tertiary structural interactions**

In collaboration with the Joe Falke Biochemistry Laboratory at the University of Colorado at Boulder, I am studying chemotaxis receptor protein sequences to identify important tertiary structural interactions that may be detectable via sequence covariation. Chemotaxis receptor proteins (specifically tar and tsr) form four-helix bundles in their cytosolic region. An amino acid residue side-chain on one helix (referred to as the knob residue) will interact with a group of amino acid residues on the adjacent helix (referred to as the hole residues) to stabilize the four helix bundle. The knob/hole structure is referred to as a socket.

Many sockets in chemotaxis receptors are perfectly conserved, but in some cases a socket is suspected, but there is not perfect residue conservation. I have identified two sockets which appear to covary, and predict that these may be structurally (and therefore functionally) important interactions. These predictions are currently being tested *in vitro* with double-mutant studies in the Falke Lab.

Finally, in collaboration with a student in the Amy Palmer Biochemistry Laboratory at the University of Colorado at Boulder, I am studying fluorescing proteins to identify pairs of positions that may have coevolved, and which therefore may be important to the tertiary structure of GFP. Identification of these position pairs will subsequently be used to make informed decisions about which positions may be interesting to modify in attempt to engineer new variants of these biotechnologically important proteins.

### **Efficient, accurate, and automated curation of biomedical data**

The quantity of many flavors of biological data is growing rapidly, including biomedical literature, fully sequenced genomes, gene expression data, individual gene and protein sequences, and molecular structures. The rate of curation of biomedical data is unable to keep pace, and a transition to partially (or eventually fully) automated annotation and database construction will be critical to extracting the most information from existing data. Biomedical text mining attempts to address the exponential growth of biomedical literature by supplementing manual curation techniques with automated approaches.

I have applied biomedical text mining techniques to the problem of identifying descriptions of protein point mutations in biomedical literature, with the goal of using the resulting system to automatically construct protein mutation databases from biomedical literature. Toward this end I have developed the highest performance point mutation extraction system currently available, MutationFinder [3], and several research groups are now using my tool. More interesting than the ability of MutationFinder to achieve very high precision and recall on blind test data, the regular expressions that drive this system were automatically generated requiring very little human interaction. This is in stark contrast to many other high-performance information extraction systems which require many person-hours of pattern generation. As specific needs arise, I am interested in expanding my automatic pattern generation approach to other information extraction tasks, and incorporating the resulting tools into database annotation tasks. I believe that a combination of manual and automated database annotation techniques can be applied to improve the coverage and accuracy of current biomedical databases, while simultaneously generating gold-standard data for

future automated annotation approaches.

## **The role of software engineering in computational biology**

Developing, testing, and publishing software is the computational biologist's primary means for conducting reproducible experiments. Documenting software is a key component of record keeping for the computational biologist, and I strongly believe that the quality and comprehensiveness of software tests is an indicator of the quality of the science.

Large scale projects such as the Human Microbiome Project, ENCODE, The 1000 Genomes Project, the Archon X-Prize for Genomics, and the many species-specific genome projects promise the rapid accumulation and (perhaps less rapid) interpretation of biological data. As biology continues to become more data-intensive the toolset of the biologist continuously expands and changes, and computational tools and statistical methods become more central. In addition to my biological research goals, I am devoted to developing high quality, highly reusable, open source bioinformatics tools based on sound software engineering principles. To date, I have made significant contributions to two bioinformatics software packages that are currently available via Sourceforge: MutationFinder [3] and PyCogent [4]. In both cases, close attention has been paid to software testing to ensure its quality, and to software documentation both at the end-user level via tutorials and examples, and at the developer-level via structured and detailed comments. I plan to continue to support these tools, and expect that my future research will result in new open source software.

## **Selected prior research projects**

### **Rob Knight Laboratory, October 2006 – March 2008**

Performed a large-scale comparison of coevolution algorithms which explicitly incorporate phylogeny (tree-aware algorithms) and those that do not (tree-ignorant algorithms). A novel method was developed for testing and comparing coevolution algorithms using protein alpha helices which has advantages over pre-existing evaluation strategies. This study illustrated that currently available tree-ignorant coevolution algorithms, which are frequently orders of magnitude faster than currently available tree-aware algorithms, can yield equivalent or better results than tree-aware methods if compared to a background distribution that implicitly controls for phylogeny. This finding is surprising, and opens the door to applications of coevolution algorithms (such as application on a genome-wide scale) which previously were thought to be too compute-intensive to be practical. This work resulted in a peer-reviewed publication in BMC Evolutionary Biology [1]; an Outstanding Research Award for my poster presentation at the University of Colorado Denver Student Research Forum in January of 2008; and approximately 5000 lines of code for performing coevolutionary analyses available as part of the open source PyCogent package.

### **Larry Hunter Laboratory, January 2007 – July 2007**

Performed an evaluation of automated approaches for database annotation. Two automatic (but very different) approaches for annotating mutations in Protein Data Bank (PDB) structures were compared with each other, and with human-annotated mutation data. The relative quality and quantity of the data generated by each of the three approaches (human and two automatic approaches) was compared to understand the best approaches for biological database annotation.

The findings indicate that none of the annotation techniques, including author annotation, are sufficient for accurate and comprehensive database annotation, and that combined manual/automatic approaches may presently be the best option in terms of quality and cost. This work resulted in a peer-reviewed publication in the Pacific Symposium on Biocomputing 2008 proceedings [5], and an oral presentation at that meeting.

#### **Larry Hunter Laboratory, December 2005 – January 2007**

Developed an automated procedure for generating biomedical information extraction systems. Focused on the problem of identifying descriptions of point mutations in free text, an automated approach was developed to create an information extraction system. The resulting point mutation extraction system achieved significant performance increases over existing systems, both in terms of precision and recall, and was developed in a small fraction of the person-hours required for other systems. This work resulted in two peer-reviewed publications [3, 6]; an open source software package, MutationFinder; and a poster presentation at the 2007 Pacific Symposium on Biocomputing.

#### **Larry Hunter Laboratory, July 2006 – November 2006**

Participated in the international 2006 Text REtrieval Conference (TREC) Genomics competitive evaluation. My work on TREC 2006 consisted of developing an automated question-answering system for biomedical research questions, using a corpus of 162,259 full-text MEDLINE articles as a pool for answers. Methods applied included semantic analysis of text, document zoning, singular value decomposition/latent semantic analysis, machine learning, and Lemur-based information retrieval. Our system produced the best result for at least one of the twenty-seven queries, in each of three evaluations. This work resulted in a TREC proceeding publication [7].

#### **Larry Hunter Laboratory, July 2005 – November 2005**

Participated in the international 2005 Text REtrieval Conference (TREC) Genomics Document Classification competitive evaluation. This project involved development of an automated document triage system to classify approximately six thousand full-text mouse biology articles into four possible categories. A system in Python, employing Support Vector Machines and Naive Bayes document classifiers, that achieved consistently high F-measures, including the highest F-measure for one subtask, and the highest precision for all tasks. This work resulted in a TREC proceeding publication [8].

#### **Robert Garcea Laboratory, August 2004 – November 2004**

Studied the effects of pH on HPV-11 viral capsid formation. Gained experience in structural virology by performing techniques including protein isolation via affinity chromatography, protein identification via SDS-PAGE and Western blotting, protein purification via FPLC, and protein structure observation via electron microscopy.

#### **Michael Yarus and Rob Knight Laboratories, September 2003 – August 2004**

Developed software in Python to investigate the roles of stereochemical and adaptive factors in the evolution of the canonical genetic code. With five different approaches for creating random

genetic codes, we compared the code error (difference in polar requirement for single-nucleotide codon interchanges) with the coding triplet concentrations in RNA binding sites for eight amino acids, and showed that these properties are independent and uncorrelated. This work led to two peer-review publications [9, 10].

## References

- [1] J. Gregory Caporaso, Sandra Smit, Brett C. Easton, Lawrence Hunter, Gavin A. Huttley, and Rob Knight. Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evol Biol*, 8(1):327, Dec 2008.
- [2] A. S. Kondrashov, S. Sunyaev, and F. A. Kondrashov. Dobzhansky-muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A*, 99(23):14878–14883, November 2002.
- [3] J. Gregory Caporaso, William A Baumgartner, David A Randolph, K. Bretonnel Cohen, and Lawrence Hunter. MutationFinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865, Jul 2007.
- [4] Rob Knight, Peter Maxwell, Amanda Birmingham, Jason Carnes, J. Gregory Caporaso, Brett C Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, Matthew J Wakefield, Jeremy Widmann, Shandy Wikman, Stephanie Wilson, Hua Ying, and Gavin A Huttley. PyCogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171, 2007.
- [5] J. Gregory Caporaso, Nita Deshpande, J. Lynn Fink, Philip E Bourne, K. Bretonnel Cohen, and Lawrence Hunter. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pac Symp Biocomput*, pages 640–651, 2008.
- [6] J. Gregory Caporaso, William A Baumgartner, David A Randolph, K. Bretonnel Cohen, and Lawrence Hunter. Rapid pattern development for concept recognition systems: Application to point mutations. *J Bioinform Comput Biol*, 5(6):1233–1259, Dec 2007.
- [7] J. Gregory Caporaso, William A. Baumgartner, Hyunmin Kim, Zhiyong Lu, Helen L. Johnson, Olga Medvedeva, Anna Lindemann, Lynne Fox, Elizabeth K. White, K. Bretonnel Cohen, and Lawrence Hunter. Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question-Answering. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.
- [8] J. Gregory Caporaso, William A. Baumgartner, K. Bretonnel Cohen, Helen L. Johnson, Jesse Paquette, and Lawrence Hunter. Concept recognition and the TREC Genomics tasks. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, 2005.
- [9] J. Gregory Caporaso, Michael Yarus, and Rob Knight. Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J Mol Evol*, 61(5):597–607, Nov 2005.
- [10] Michael Yarus, Gregory J. Caporaso, and Rob Knight. Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem*, 74:179–198, 2005.