

Exploring Genomic Context Patterns for *Rhodobacter sphaeroides* in the HERBE Knowledge Discovery Environment

Heidi J. Sofia, Abigail L. Corrigan, Kyle R. Licker, George Chin, and Eric G. Stephan
Pacific Northwest National Laboratory, 902 Battelle Blvd., Richland, WA 99352, USA
{Heidi.Sofia, Abigail.Corrigan, Kyle.Licker, George.Chin, Eric.Stephan}@pnl.gov

Abstract

Sophisticated information strategies are increasingly essential for biologists. We have built a powerful knowledge discovery resource using novel genomic context methods, interactive visualization strategies, and computational environment technologies. The Heuristic Entity Relationship Building Environment (HERBE) is a research platform for advanced database technologies that fuses data management solutions with knowledge management components to support the dynamic capture of concepts and observations as biologists explore large-scale data. The Similarity Box visualization software supports the ability of biologists to interact with large-scale computational results and evaluate relationships based on natural reasoning processes. We have applied these knowledge methods in the exploration of complex microbial genome relationships. We extracted a complete set of gene neighbor patterns for *Rhodobacter sphaeroides* using HERBE to map data structures for chromosomal contiguity against sequence similarity. We then organized these gene neighbor patterns by their phylogenetic profiles using Similarity Box to enable biologists to explore the results.

1. Introduction

New "genomic context" methods are being developed to extract higher-order relationships such as phylogenetic profiles and conserved gene neighbor patterns from large sequence databases [1]. These approaches combine simpler data types to describe more complex patterns in the context of many completed genomes. Phylogenetic profiles are the fingerprint of occurrence for a feature among different species. For example, certain proteins are strong markers for particular biochemical pathways, and these can be unique indicators for organisms that make a

"living" in certain ways, for example, light-harvesting proteins in photosynthetic species. Gene neighbor methods detect conserved clusters of genes that are closely adjacent in prokaryotic chromosomes across multiple species. Chromosomal rearrangements cause the order of genes to be randomized much like a deck of cards that is shuffled. However, in prokaryotes, genes are sometimes "sticky" and resist this process. Conserved gene neighbors provide an important clue because their gene products often interact in complexes or are part of the same pathway.

Database strategies can also be employed to organize data into new structures that represent knowledge to biologists. However the capabilities of a database are predetermined by the assumptions and technologies inherent within the database design, data models, and query mechanisms. Computer scientists who build data models are essentially capturing scientific knowledge in these representations and structures [2]. However, building explicit data models in biology is difficult because of complexity, uncertainty, and rapid change of the information.

2. Advanced knowledge technologies

The Heuristic Entity Relationship Building Environment (HERBE) is a prototypical architecture that is implemented in Java and PostgreSQL to fuse data management technology with knowledge management components [3]. This hybrid approach enables biologists to manage both concepts and related data in a synchronized fashion. Import mechanisms have been built for analysis results from BLAST [4], Similarity Box [5], and other sources to extract concepts into uniform data structures, which are stored in a Concept Repository. Data and data descriptions are used to evolve an underlying database.

The Similarity Box software is a Java genome comparison platform that places a dynamic layer over BLAST and GenBank resources using interactive

visualizations such as dendrogram, network, and linked set views [5]. This platform is designed to add genomic context analysis to the standard bioinformatics toolkit available to biologists, which currently includes similarity and motif searches, multiple alignment, phylogenetic analysis, and sequence feature extraction.

3. Genomic context analysis

Protein and similarity relationships were entered into HERBE for *R. sphaeroides* using a draft version of 3928 predicted open reading frames (ORFs) for the two chromosomes, not including plasmid proteins. Each protein sequence was used as a BLAST query against the nr database to define similarity sets, for a total of 60,616 proteins. Of these, 47097 prokaryotic sequences were identified, retrieved in FASTA format, and entered into HERBE. Nucleotide records for these sequences were used as input in the Neighbor Finder function of the Similarity Box software to extract proteins encoded as adjacent genes from these records. These results were entered as a gene neighbor association table. HERBE gene neighbor queries are reported as a Web table as shown in Figure 1.

Complete gene neighbor patterns for the *R. sphaeroides* genome were exported from the HERBE database into a matrix which was imported into Similarity Box for hierarchical clustering. This organized view of genome-scale conserved gene neighbor patterns guides the biologist in connecting local gene clusters into higher-order patterns.

RSP0091	RSP0092	RSP0090
1552333	2262748	2262747
1552332	2262748	2262747
1552331	2262748	2262747
1552330	2262748	2262747
1552329	2262748	2262747
1552328	2262748	2262747
1552327	2262748	2262747
1552326	2262748	2262747
1552325	2262748	2262747
1552324	2262748	2262747
1552323	2262748	2262747
1552322	2262748	2262747
1552321	2262748	2262747
1552320	2262748	2262747
1552319	2262748	2262747
1552318	2262748	2262747
1552317	2262748	2262747
1552316	2262748	2262747
1552315	2262748	2262747
1552314	2262748	2262747
1552313	2262748	2262747
1552312	2262748	2262747
1552311	2262748	2262747
1552310	2262748	2262747
1552309	2262748	2262747
1552308	2262748	2262747
1552307	2262748	2262747
1552306	2262748	2262747
1552305	2262748	2262747
1552304	2262748	2262747
1552303	2262748	2262747
1552302	2262748	2262747
1552301	2262748	2262747
1552300	2262748	2262747
1552299	2262748	2262747
1552298	2262748	2262747
1552297	2262748	2262747
1552296	2262748	2262747
1552295	2262748	2262747
1552294	2262748	2262747
1552293	2262748	2262747
1552292	2262748	2262747
1552291	2262748	2262747
1552290	2262748	2262747
1552289	2262748	2262747
1552288	2262748	2262747
1552287	2262748	2262747
1552286	2262748	2262747
1552285	2262748	2262747
1552284	2262748	2262747
1552283	2262748	2262747
1552282	2262748	2262747
1552281	2262748	2262747
1552280	2262748	2262747
1552279	2262748	2262747
1552278	2262748	2262747
1552277	2262748	2262747
1552276	2262748	2262747
1552275	2262748	2262747
1552274	2262748	2262747
1552273	2262748	2262747
1552272	2262748	2262747
1552271	2262748	2262747
1552270	2262748	2262747
1552269	2262748	2262747
1552268	2262748	2262747
1552267	2262748	2262747
1552266	2262748	2262747
1552265	2262748	2262747
1552264	2262748	2262747
1552263	2262748	2262747
1552262	2262748	2262747
1552261	2262748	2262747
1552260	2262748	2262747
1552259	2262748	2262747
1552258	2262748	2262747
1552257	2262748	2262747
1552256	2262748	2262747
1552255	2262748	2262747
1552254	2262748	2262747
1552253	2262748	2262747
1552252	2262748	2262747
1552251	2262748	2262747
1552250	2262748	2262747
1552249	2262748	2262747
1552248	2262748	2262747
1552247	2262748	2262747
1552246	2262748	2262747
1552245	2262748	2262747
1552244	2262748	2262747
1552243	2262748	2262747
1552242	2262748	2262747
1552241	2262748	2262747
1552240	2262748	2262747
1552239	2262748	2262747
1552238	2262748	2262747
1552237	2262748	2262747
1552236	2262748	2262747
1552235	2262748	2262747
1552234	2262748	2262747
1552233	2262748	2262747
1552232	2262748	2262747
1552231	2262748	2262747
1552230	2262748	2262747
1552229	2262748	2262747
1552228	2262748	2262747
1552227	2262748	2262747
1552226	2262748	2262747
1552225	2262748	2262747
1552224	2262748	2262747
1552223	2262748	2262747
1552222	2262748	2262747
1552221	2262748	2262747
1552220	2262748	2262747
1552219	2262748	2262747
1552218	2262748	2262747
1552217	2262748	2262747
1552216	2262748	2262747
1552215	2262748	2262747
1552214	2262748	2262747
1552213	2262748	2262747
1552212	2262748	2262747
1552211	2262748	2262747
1552210	2262748	2262747
1552209	2262748	2262747
1552208	2262748	2262747
1552207	2262748	2262747
1552206	2262748	2262747
1552205	2262748	2262747
1552204	2262748	2262747
1552203	2262748	2262747
1552202	2262748	2262747
1552201	2262748	2262747
1552200	2262748	2262747
1552199	2262748	2262747
1552198	2262748	2262747
1552197	2262748	2262747
1552196	2262748	2262747
1552195	2262748	2262747
1552194	2262748	2262747
1552193	2262748	2262747
1552192	2262748	2262747
1552191	2262748	2262747
1552190	2262748	2262747
1552189	2262748	2262747
1552188	2262748	2262747
1552187	2262748	2262747
1552186	2262748	2262747
1552185	2262748	2262747
1552184	2262748	2262747
1552183	2262748	2262747
1552182	2262748	2262747
1552181	2262748	2262747
1552180	2262748	2262747
1552179	2262748	2262747
1552178	2262748	2262747
1552177	2262748	2262747
1552176	2262748	2262747
1552175	2262748	2262747
1552174	2262748	2262747
1552173	2262748	2262747
1552172	2262748	2262747
1552171	2262748	2262747
1552170	2262748	2262747
1552169	2262748	2262747
1552168	2262748	2262747
1552167	2262748	2262747
1552166	2262748	2262747
1552165	2262748	2262747
1552164	2262748	2262747
1552163	2262748	2262747
1552162	2262748	2262747
1552161	2262748	2262747
1552160	2262748	2262747
1552159	2262748	2262747
1552158	2262748	2262747
1552157	2262748	2262747
1552156	2262748	2262747
1552155	2262748	2262747
1552154	2262748	2262747
1552153	2262748	2262747
1552152	2262748	2262747
1552151	2262748	2262747
1552150	2262748	2262747
1552149	2262748	2262747
1552148	2262748	2262747
1552147	2262748	2262747
1552146	2262748	2262747
1552145	2262748	2262747
1552144	2262748	2262747
1552143	2262748	2262747
1552142	2262748	2262747
1552141	2262748	2262747
1552140	2262748	2262747
1552139	2262748	2262747
1552138	2262748	2262747
1552137	2262748	2262747
1552136	2262748	2262747
1552135	2262748	2262747
1552134	2262748	2262747
1552133	2262748	2262747
1552132	2262748	2262747
1552131	2262748	2262747
1552130	2262748	2262747
1552129	2262748	2262747
1552128	2262748	2262747
1552127	2262748	2262747
1552126	2262748	2262747
1552125	2262748	2262747
1552124	2262748	2262747
1552123	2262748	2262747
1552122	2262748	2262747
1552121	2262748	2262747
1552120	2262748	2262747
1552119	2262748	2262747
1552118	2262748	2262747
1552117	2262748	2262747
1552116	2262748	2262747
1552115	2262748	2262747
1552114	2262748	2262747
1552113	2262748	2262747
1552112	2262748	2262747
1552111	2262748	2262747
1552110	2262748	2262747
1552109	2262748	2262747
1552108	2262748	2262747
1552107	2262748	2262747
1552106	2262748	2262747
1552105	2262748	2262747
1552104	2262748	2262747
1552103	2262748	2262747
1552102	2262748	2262747
1552101	2262748	2262747
1552100	2262748	2262747
1552099	2262748	2262747
1552098	2262748	2262747
1552097	2262748	2262747
1552096	2262748	2262747
1552095	2262748	2262747
1552094	2262748	2262747
1552093	2262748	2262747
1552092	2262748	2262747
1552091	2262748	2262747
1552090	2262748	2262747
1552089	2262748	2262747
1552088	2262748	2262747
1552087	2262748	2262747
1552086	2262748	2262747
1552085	2262748	2262747
1552084	2262748	2262747
1552083	2262748	2262747
1552082	2262748	2262747
1552081	2262748	2262747
1552080	2262748	2262747
1552079	2262748	2262747
1552078	2262748	2262747
1552		