A Bioinformatics Software Company

Bio::Neos, Inc.
100 Oakdale Campus, TIC #1
Iowa City, IA 52242 USA
Telephone: 319.248.9610
Facsimile: 319.335.4489

Memo of Understanding

VCMap Redesign Project

Prepared for: Dr. Anne Kwitek

***Disclaimer**: This document is meant to serve as a summarized description of a proposed project between Dr. Kwitek and Bio::Neos and cannot be used as a legally binding contract, official or accurate cost or time estimate, or official statement of work.*

I. Project Overview

Dr. Kwitek's research relies on comparative mapping and comparative data mining. Currently, Dr. Kwitek frequently utilizes the Rat Genome Database, hosted by the Medical College of Wisconsin (MCW), and the public VCMap tool. This tool is outdated, inefficient and under-supported. Dr. Kwitek is interested in working together with Bio::Neos and key members at the MCW in order to facilitate a redesign of this tool or the design of a successor to this tool. The funding for this project will primarily come from Dr. Kwitek, but there may be additional funding available from other sources if needed.

II. Proposed Solution

Bio::Neos is interested in working with Dr. Kwitek and any of her collaborators in order to develop an updated version of the VCMap tool to facilitate more efficient comparative genomic research. The initial solution will provide most, or all, of the current features of the VCMap software. The tool will support Human, Rat, and Mouse genomes initially, but the application will be built in a modular way that supports adding additional genomes without major design changes. Unlike the existing application, the new solution will have either a web based, or web accessible front-end interface that communicates with a database back-end to dynamically load and unload data without requiring the users to restart the application.

1. Critical Features

It is absolutely necessary for the new tool to implement these features in order for it to meet the needs of Dr. Kwitek and her colleagues. These features will be considered throughout the design process.

i. Comparative Maps

This is the main data entry point for users of the VCMap tool. These “virtual maps” are used to demonstrate relationships between different genomes according to homology. In VCMap, the comparative map for one species is designated as the backbone onto which all annotation and other maps are anchored. The new application will likely follow this same paradigm for loading information.

ii. **QTL Annotation**
Annotated Quantitative Trait Loci (QTLs) are one of the most important pieces of annotation used by Dr. Kwitek in her research. These loci are used in order to determine regions of interest for future experiments. Thus, QTL data must be supported by the new version of VCMap. This data will be retrieved from several public data sources, including the rat genome project at MCW.

iii. **Modularity**
The new tool must be built in a modular manner that will facilitate adding new components to the software. These components could be additional genomic data, new types of annotation, different ways to display annotation, or other components.

Even though the tool will be as modular as possible, it is understood that certain unanticipated features could require redesigning parts of the application. Our goal is simply to minimize the amount of the application affected by that situation.

iv. **Dynamic Data Access**
The existing VCMap software was designed to operate on a series of flat files created by external scripts. In order to access additional data after loading the application, users are forced to start over and rerun the scripts and reload the applet.

This inefficient work flow will be eliminated in the new application by supporting dynamic data access. This will be achieved by designing a database to store the necessary data and allowing the front-end interface to communicate with this database dynamically. The design pattern for this database will be one of either a data warehouse that mirrors subsets of data from publicly available databases, or a minimal system that stores access identifiers for other public databases, allowing for real-time data aggregation.

v. **Web Accessibility**
The new system will need to be accessible via a modern web browser in order to facilitate access for any researchers interested in utilizing the new application and promote widespread usage of the software. The current VCMap, despite its significant drawbacks, has a significant user base and any replacement solution must support that existing user base. Additionally, it is expected that the solution will provide links to outside sources of publicly available information accessible on the web. Because the new solution will be designed in a web-based environment, this should not pose a problem.

2. **Optional Features**

These features could potentially increase the usefulness of the new software tool, but they are not necessary in the initial version of the software tool. If cost and time budgets allow, these features will be implemented, but if not, they can be added at a later time. Initial software design should consider these features in order to avoid any potential design pitfalls that would hinder the addition of these features in future versions of the software.

i. **Genome Maps**
Comparative data maps often do not supply enough information to anchor annotation from other data sources. In order to anchor pieces of annotation onto these "virtual" maps, often a genomic map is needed as an intermediate step to translate the coordinates between maps. For those species that have completed genomic maps, it is very useful to be able to load genomic maps, along with the comparative maps, in order to correctly orient as much annotation to the backbone sequence as possible.

ii. **RH Maps**
For species that do not have a completed genomic map, there are alternative ways to translate annotation coordinates between maps. Often this involves Radiation Hybrid (RH) maps. In certain instances these maps are even useful for loading annotation for species that have completed genomic maps. It will be useful to be able to load these maps, along with the comparative maps, in order to anchor more annotation to the backbone sequence.

iii. **Support for additional annotation**
Additional annotation also provides informative details about genomic regions of interest. This includes, but is not limited to, Copy Number Variations (CNVs), Single Nucleotide Polymorphisms (SNPs), Simple Sequence Length Polymorphisms (SSLPs), Sequence Tag Sites (STSs), and others. It would be useful to display as much annotation as possible for the users of the application to examine.

iv. **Non-sequence based annotation**
All of the maps and annotation described up to this point are related to sequence data. Additional annotation that is not related to sequence data may be useful for researchers that are using this tool. These annotations potentially include both qualitative and quantitative data. For example, expression within a particular tissue could be qualitative (present/absent) or quantitative (RMA summarized value).

3. **Database**
As mentioned above, in order to efficiently support dynamic data access throughout the application and maintain relationships between pieces of annotation, it will be necessary to utilize a relational database on the back-end of the system. We will be able to leverage existing mature open-source solutions, such as MySQL, for this application in order to reduce costs. The

database design will be an important and involved portion of the design for this application. The location where the final database for this project will be hosted is not yet determined, although development versions of the database will be hosted by Bio::Neos.

III. Cost Estimate

The cost of this project is not yet determined. Bio::Neos will make every effort to suggest a reasonable development path that will meet the needs of Dr. Kwitek while staying within her budget. This means that features will be prioritized and less critical features may potentially be delayed in order to stay in budget. As details of this project are clarified, Bio::Neos will provide an accurate estimate of the cost of the project prior to the commencement of any development work. Currently the projected estimate for this project is three to five months and $25,000 to $35,000.

IV. Alternative Solutions

1. **Passive solution**
   Potentially, a modern, modular, efficient system without a database back-end could be designed. This system could either rely on flat files and existing back-end scripts, or new scripts that dynamically create flat files. In either case, a relational database would not be used to support the back-end of the system. This solution would not be as dynamic as possible, but with a solid modular design, the interface could be adapted to support a database back-end in the future. This would potentially reduce the time and cost of the initial application, while allowing for additional updates in the future.
2. **Update existing VCMap**
   Although the existing solution is rather outdated and may suffer from a lack of modularity in its core design, it may be possible to update the existing software to support a dynamic, database driven back-end. This will require coordination between Bio::Neos and the existing support staff for the VCMap software, access to the original VCMap and supporting script source code, and possibly access to the servers on which this software is currently stored. This solution would be more administratively complex, and it is unclear if it would result in a less expensive solution. The resulting software solution would likely not meet all of the goals defined above.