Summary of Responses to CORBAmed RFI 4 (Revision 1)

Life Sciences Research DSIG White Paper

OMG Document number 98-01-01

This document contains a summary of responses submitted to the Life Sciences Research (LSR) RFI (OMG Document #corbamed/97-09-16) prior to the December 1997, OMG Technical Meeting. It was prepared by the LSR RFI Review Working Group. This document will be used as a reference to the information received by the LSR DSIG as a result of the above RFI and will act as a guide for future actions by the LSR group.

1 Overall Summary

As listed in Table 1, there were 15 submissions. The submissions were first classified according to the following vertical subdomains.

- Structural biology - **S/Bio**
- Genomics - **G**
- Cheminformatics - **C**
- Bioinformatics - **B**

Table 1 An X in the column indicates that the submission contains discussion pertinent to the indicated category as given in Section 2 of the RFI. A - in the column indicates that the submission was not evaluated with respect to these categories

Submitting Organization	Submission ID	Architecture	Inter-operabil.	Object/Data Model	Interfaces	Existing Impl.	Standards	Legal Issues	Sub-Domain
Birkbeck College	07	-	-	-	-	-	-	-	S/Bio
Genome Database	08/09			X	X	X			G
Oxford Molecular Group	10				X	X	X		B,C
Roslin Institute	11					X			G
Univserity of Manchester	12				X	X			-
University of London	13		X		X	X			B
National Center for Genome Resources	14					X			B
Sequana Therapeutics	15	X		X	X	X			B,G
Bioperl Developers	16			X	X	X			G,S/Bio
Tripos	17	-	-	-	-	-	-	-	C
NetGenics	18	X	X		X	X	X		B,G
European Bioinformatics Institute	19	X	X	X	X	X	X		G,S/Bio
Berkeley Drosophila Genome Center	20					X			G
Infobiogen	21					X			G
University of Pennsylvania	22	X	X	X	X	X			B,G

The submission of the University of Manchester presents a general system for accessing heterogeneous information resources and did not fall specifically into any of the subdomains listed in Table 1. The Tripos submission was the only substantial submission in the area of cheminformatics. The submissions from Birkbeck College and the European Bioinformatics Institute were the only submissions in the area of structure biology. Because of the low representation in these subdomains it was decided to defer examination of these responses[1] until a more representative set becomes available. The remaining submissions (including the University of Manchester's) were additionally classified by the information presented with respect to the categories given in Section 2 of the RFI.

[1] The portions of the EBI response not relating to structure biology were examined in more detail.

2 Architecture and Interoperability

In Table 2, submissions that showed special attention to architecture and interoperability are further categorized. The Netgenics response was the only one that presented a software architecture. The others presented discussions on issues relevant to such architectures. The Netgenics architecture is a general purpose architecture that has been specifically populated with objects relevant to the life sciences research domain. Much of the discussion in this submission is general to software architectures outside of life sciences. The Sequana Therapeutics and European Bioinformatics Institute responses provide lists of services and components of specific interest to bioinformatics and genomics. The University of Pennsylvania response presents architectural issues in these areas from the point of view of Graphic User Interfaces (GUIs).

Table 2

Submission	2.1.1.1Architecture Type	**Architecture Components Required**	**Interoperability Level**
European Bioinformatics Institute (#19)	General framework for managing and interacting with biological information. Components were listed; inter-component interaction diagrams/models missing.	Persistence Data transfer Analysis Curation/annotation Visualization Query Data collection/harveting	Data Inter-component
University of Pennsylvania (#22)	Java-based client-side Visualization framework for genomics data	Sequences Generalized sequence annotation Maps BLAST results Multiple sequence alignments Map alignments Chromosomes	MVC-like layer decoupling data from views and supporting synchronization of feature selection among views
Sequana Therapeutics (#15)	Specialized genomics framework	Data visualization Computational resource management Analysis service Instrumentation	Data resources
Netgenics (#18)	General purpose business object architecture (project management, groupware, application management) with object interaction diagrams.	General purpose components listed. Precondition for viability of architecture is well-defined domain object model(s).	CORBA

3 Object Data Models and Interfaces

The following submissions made explicit mention (and contained descriptions) of domain objects for which object models and/or interfaces are desired:

- European Bioinformatics Institute
- Sequana Therapeutics
- Bioperl Developers
- University of Pennsylvania

Table 3

Sequencing projects	**Sequences/features**	**Map**	**Chromosome**	**Structure**	**Data Resources**	**Analysis tools**
Clone	Genes	Interval	Collection of genes	3D Structure	DNA construction (sequence assembly)	BLAST
Template	Exons	Radiation	Telomere		Alignment data (and scoring matrices)	Multi-sequence alignment
Read		Linkage	Centromere		Pattern and profile (motif searching)	Sequence pattern search
Assembly/contig/consensus		Cytogenic			Phylogeny and taxonomy	Restriction enzyme
Super contig/super assembly/contig map					Biochemical pathways	
					References to bibliographic materials	

These object types are summarized in Table 3, where they have been classified into 6 categories:

1. Objects of interest to DNA sequencing projects
2. Map objects
3. Chromosome objects
4. Three-dimensional structure objects
5. Data resource objects
6. Analysis tools

The EBI response suggested that we distinguish objects contained in data repositories from objects that operate on this underlying information as reflected in categories 5 and 6. Table 4 gives a complete list of all object classes (and subclasses derived from them) as listed in the various submissions. Table 5 shows details of interfaces to databases and analysis tools described in the submissions.

Table 4

Name of Class	Derivative	Alternate Name	Source	Description
algorithm			Bioperl	Setting parameters, generating, and executing commands for computational analyses
allele			Sequana	A specific variant of at a variant site
annotation		annotation (I)	U Penns	NEED MORE INFO
	pointAnnotation (I)		U Penns	NEED MORE INFO
	spanAnnotation (I)		U Penns	NEED MORE INFO
	functional		Bioperl	Container for functional information about a gene product
assembly			Sequana	The contigs computed by an assembly engine using all reads for a sequence project as input
centromere			Bioperl	Represents a chrmosome's centromeric sequences
chromosome			Bioperl	Generic chromosome object; container for gene objects
clone				
	cDNA clone		Sequana	A DNA copy of all or part of an RNA molecule (usually mRNA)
	genomic clone		Sequana	A clone of human genome DNA from a BAC, PAC, or cosmid library
consensus			Sequana	The most likely sequence given a layout (assembly) of reads
contig			Sequana	A consensus sequence and mapping of reads on the consensus
domain			Bioperl	Represents a protein domain (derives from Bio::Struct::Domain)
EST			Sequana	A partial sequence from a cDNA clone
			EBI	Object model specification for EST clone library, clone object, and sequence
exon			Sequana	
			Bioperl	Represents a single exon within a gene
expression pattern			Sequana	A profile describing the tissue and cell type pattern of expression of a gene
feature			Sequana	Information about a doman object
gene			Bioperl	Generic gene object; container/factory for protein objects
genome			Bioperl	Genomic information manager and factory for gene, protein, and chromosome objects
			EBI	Data model to capture structural and functional for bacteria at transcriptional, translational, and regulatory levels
genotype			Sequana	A set of specific alleles for an individual
haplotype			Sequana	A set of alleles for a given chromosome
intron			Sequana	NEED MORE INFO
			Bioperl	Represents a single intron within a gene

Name of Class	Derivative	Alternate Name	Source	Description
map				
	feature map		Sequana	An interval map of features relative to an object's coordinate system (I.e., features on a sequence)
	cytogenic map		Sequana	An ordered map of cytogenic bands representing a chromosome
	radiation hybrid map		Sequana	A type of map ordering markers on a chromosome
			EBI	Object model for RH maps
	ordered map		Sequana	An ordering of objects
	interval map		Sequana	A mapping of objects to their positions relative to some coordinate system
	genome		EBI	Object model for genome maps
mRNA			Sequana	NEED MORE INFO
open reading frame		ORF	Bioperl	Represents open reading frame (protein-coding) regions of gene objects
ortholog			Sequana	Genes or proteins from two different species that have a common ancestor and perform the homologous function in the organisms
paralog			Sequana	A member of a gene family
protein			Bioperl	Generic protein object; container/factory for domain objects
			EBI	Data model to represent protein sequences
read			Sequana	The sequence data from one sequencing reaction on a template
	forward read		Sequana	Sequence data generated from the 'forward' primer on a template
	reverse read		Sequana	Sequence data generated from the 'reverse' primer on a template
read layout			Sequana	An interval map of sequencing reads relative to a consensus sequence
restriction enzyme			Bioperl	Segments DNA sequences based on the presence of known restriction enzyme recognition sites
sequence			Sequana	NEED MORE INFO
	nucleotide sequence		Sequana	NEED MORE INFO
	protein sequence		Sequana	NEED MORE INFO
	sequence (I)		U Penns	Extends interval (I) I NEED MORE INFO
		sequence	Bioperl	Represents a single nucleic acid or amino acid sequence
		sequence (several)	Bioperl	SEVERAL SCOPED VARIANTS
sequence alignment				
	multiple seq. align.	sequenceGap (I)	U Penns	NEED MORE INFO
	multiple seq. align.	sequenceAlignment	U Penns	Extends interval (I) I NEED MORE INFO
		sequence alignment	Bioperl	Manipulation and analysis of multiple sequence alignments
sequence pattern			Bioperl	Creates and manipulates regular expression sequence patterns
sequencing project			Sequana	All data associated with determining the sequence of a genomic clone
similarity			Sequana	Regions of two sequences that are similar and characterized by a similarity score
structure				

Name of Class	Derivative	Alternate Name	Source	Description
	3D structure		Bioperl	Represent 3D macromolecular structures as well as core data-management and analysis features
	Chain		Bioperl	Represents data for a single chain within a 3D structure
	residue		Bioperl	Represents data for a single residue in a chain within a 3D structure
	macromol. structure		EBI	ER model to capture macromolecular structure based on PDB
supercontig			Sequana	An ordered and oriented linking of contigs
telomere			Bioperl	Represents a chromosome's telomeric sequence
template (sequencing)			Sequana	A subclone of a genomic clone in a sequencing vector
tiling path			Sequana	An interval map of clones covering some region of DNA
untranslated region				
	5' untranslated region		Bioperl	Represents 5' (upstream) untranslated region of a gene object
	3' untranslated region		Bioperl	Represents 3' (downstream) untranslated region of a gene
variant site		polymorphic site	Sequana	A position in a genome that is variable in a population of individuals

Table 5. Interfaces to data and analysis resources

BLAST Report	Bioperl	Generating, parsing, and analyzing BLAST sequence analysis reports
Blast	Bioperl	Parsing and analysis of BLAST results
Struct	Bioperl	Represents a single 3D molecuar structure and core analysis tools
PDB	Bioperl	PDB database information accessor and PDB object factory
Scop	Bioperl	SCOP database information accessor and SCOP object factory
EMBL	EBI	Interface to EMBL Nucleotide Sequence DB (includes GenBank and DDBJ)
SWISS-PROT	EBI	Interface to SWISS-PROT
EMBL Alignment	EBI	Interface to the EBI alignment database
Bacterial Genome DB	EBI	Interface to bacterial genome DB (includes MICADO?)
MSD	EBI	Interface to macromolecular structure DB
Rhdb	EBI	Interface to radiation hybrid map DB
IARC P53	EBI	Interface to mutation detection DB
dbEST	EBI	Interface to ancillary EST DB
UniGene	EBI	Interface to human gene cluster DB
THCs	EBI	Interface to tentative human consensus sequence DB
STACK	EBI	Interface to human alignment and consensus DB
BioImage	EBI	Interface to multidimensional biological images DB
PRINTS	EBI	Interface to protein motif fingerprint DB
EMP	EBI	Interface to metabolic pathways DB

Archive linking services	EBI	Interface to link various data sources
SRS	EBI	Interface to the indexing and cross-referencing services of the Sequence Retrieval System

4 Summary of Existing Implementations

The RFI requests information on availability, maturity and importance of existing object-oriented and/or legacy implementations that are being applied to life sciences research problems. Most responses to the RFI mention applications that the submitters consider significant. In some responses a extensive description is provided. Importance is never explicitly rated, but the fact that these implementations are mentioned signifies that they are important.

Table 6 summarizes the implementations mentioned in the RFI responses. A short description, extracted from the RFI response, is provided.

Table 6 Existing Implementations

Doc #	Organisation	Existing Implementations
08/09	Genome Database	OO schema of genes, PCR primers, (STSs), clones and maps
10	Oxford Molecular Group	Legacy apps including ▪ GCG Wisconsin Package a Comprehensive suite of seq analysis tools ▪ PC based sequence analysis packages GCG config files proposed as a standard & presentation given on same
11	Roslin Institute	CORBA-enabled implementation of genome mapping application - Anubis Genome viewer/browser (Java) Preliminary IDL and object model Summary information on Object and Data Model available
12	University of Manchester	Knowledge base, TAMBIS: client - General IDL requirements provided.
13	University College London	PRINTS - motif data collection CINEMA multiple sequence alignment editor client
14	National Center Genome Resourses	C++ classes for sequence analysis
15	Sequana Therapeutics	Proprietary data model, implementation and framework
16	BioPerl Developers	Sequence and alignment objects
18	NetGenics	Synergy as a standard framework architecture
19	European Bioinformatics Institute	Databases ▪ DNA ▪ Proteins ▪ Mutations ▪ Motifs ▪ Expression ▪ Pathways ▪ Polymorphisms ▪ Mapping data SRS CORBA implementations include ▪ Wrappers to RHdb, ESTs and EMBL DBs

Doc #	Organisation	Existing Implementations
		▪ Further implementations in development. ▪ Applab -generic CORBA wrapper generator for command line apps (related to GCG config files). ▪ EST alignment engine and CORBA interface.
20	Berkeley Drosophila Genome Center	CORBA enabled implementation of client-server database for fly genomic data
21	Infobiogen	Proprietary OO DBMS - CORBA interface under development
22	University of Pennsylvania	Clients (Viewing tools) that would have IDL requirements

4.1 Groups

Using this summary of the responses to RFI 1 the groups outlined below were identified. Further groups can be identified (e.g. example legacy applications which can help to identify useful objects). Further details from some of the bodies making responses might allow augmentation of these groups.

4.1.1 Client applications:

Applications that would benefit from access to CORBA based services. The listed responses specify needs for certain functionality of these services. Future IDL interface specifications could take these requirements into account.

- 11 Roslin Institute
- 12 Manchester
- 21 Infobiogen
- 22 U Penn

4.1.2 Example implementations

Applications based on OO technology, which demonstrate possible object/data models.

- 15 Sequana
- 16 BioPerl
- 18 NetGenics
- 19 EBI
- 20 Berkeley Drosophila Genome Center

4.1.3 CORBA enabled implementations

Applications that incorporate IDL interfaces and are accessible via a CORBA compliant ORB.

- 11 Roslin Institute
- 18 NetGenics
- 19 EBI
- 20 Berkeley Drosophila Genome Center

5 Standards

Many responses list existing standards, either de-facto or published by a standards body, that are considered relevant for the domain of life sciences research.

A list of the standards mentioned in the submissions is given below.

- ASCII
- Feature Table - EMBL/DDBJ/GenBank
- IUPAC residue representations (AA & NA)
- EMBL flat file format
- DDBJ & GenBank flat file format
- PIR/PIR-codata
- PDB flat file format
- SwissProt flat file format
- NCBI ASN1
- GCG
- FASTA
- Alignment formats
  - GCG-msf
  - Phylip
  - etc.
- Enzyme
- Prosite
- ?Rebase
- csf - trace data