Microsoft® Research
Faculty Summit 2010

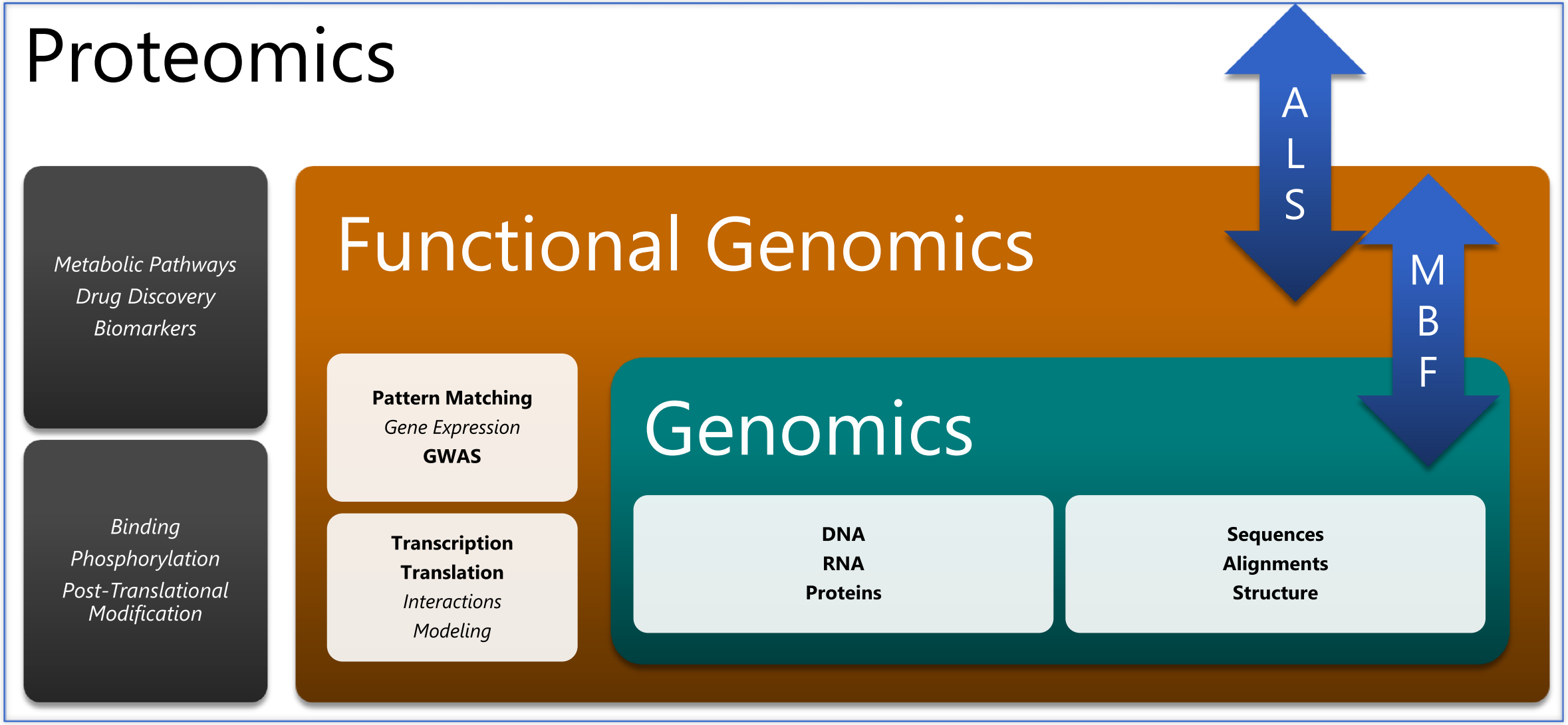
The Microsoft Biology Foundation

Simon Mercer
Director, Health & Wellbeing
Microsoft Corporation

Vision

An open-source library of reusable bioinformatics algorithms and functions built on the .NET platform

Domain and Focus of MBF



Challenges and Scenarios

Customer Challenges

- Dependency on error-prone script-based pipelines
- Challenged by increasing data sizes
- Hard to preserve tribal knowledge
- *Scientists as programmers – issues of architecture and maintainability*

Target Scenarios

- Rapid development of new applications for scientists
- Support for 'commodity' genomic sequencing
- Support for cross-species comparisons (comparative genomics)
- Support for genomic alignment manipulation
- Single-Nucleotide Polymorphism (SNP) identification

Value Proposition

Developer (Bioinformatician)

- Reduce time to develop solutions in the Genomics space
- Leverage other Microsoft technologies in Genomics applications
- Free, open source, broad license, adaptable

End-user (Biologist)

- Decrease time to insight…faster research
- Increased value of the desktop in genomics research
- Value of community collaboration

Potential for Tech Transfer

- Health Solutions Group
  - Linkage with the Amalga Life Science platform for discovery informatics
- Technical Computing
  - Integration with the Scientific Workbench
- HPC
  - Cluster-based solutions for genomics research
- Azure
  - Cloud-based solutions for genomics research

Microsoft® Research
Faculty Summit 2010

Version 1 of the Microsoft Biology Foundation has been released!

announcing

What is MBF?

- Microsoft Biology Foundation (MBF) is a bioinformatics toolkit
  - built on top of the .NET Framework 4.0
  - open source under MS-PL license
  - foundation upon which other tools can be built
- Provides various components useful for biological analysis
  - parsers to read and write common bioinformatics formats
  - support for DNA, RNA and protein sequences
  - algorithm framework for analysis and transformation
  - web connector framework for web-service interaction

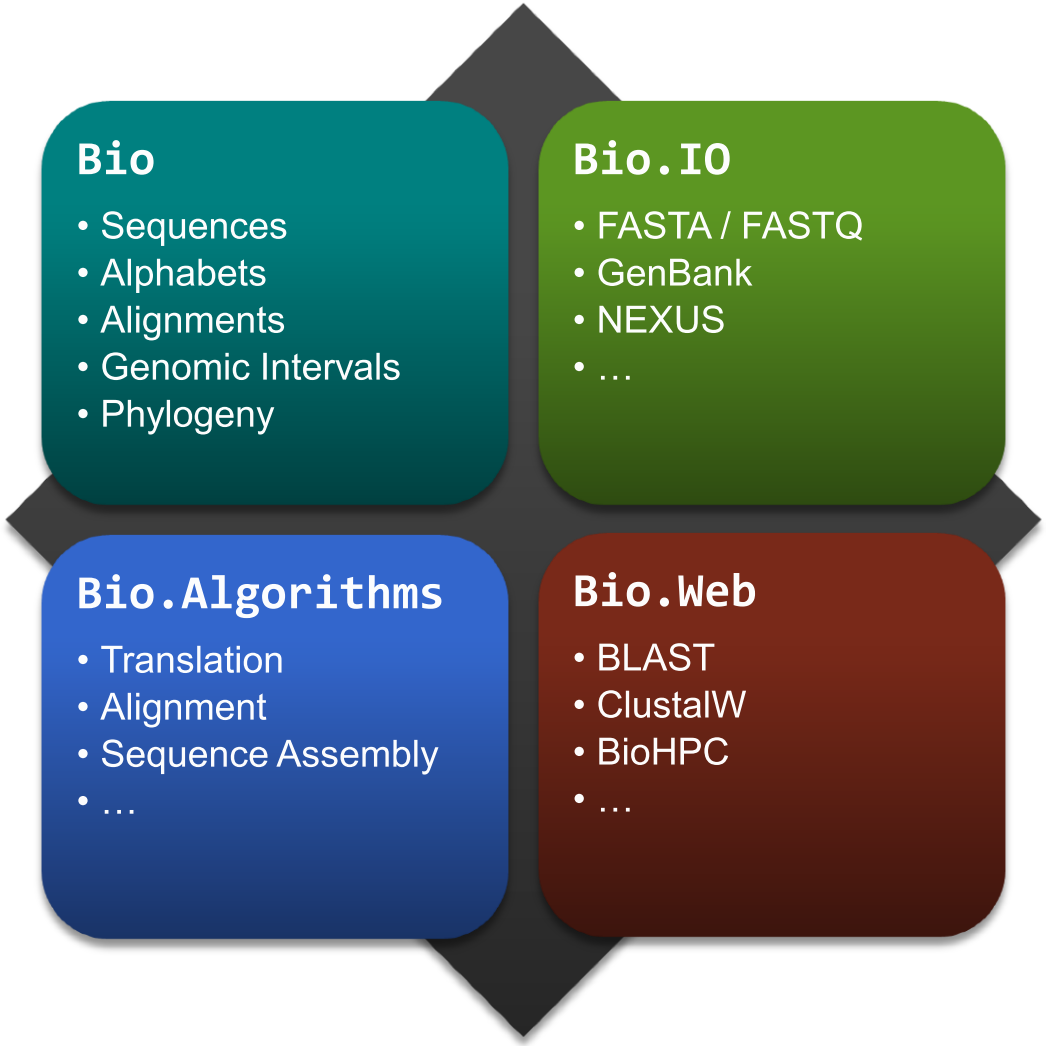
What is MBF intended to do?

- Primarily focused on genomics
  - reusable data structures to represent sequences + symbols
  - I/O framework to load/save sequences
  - algorithm framework to process loaded sequences
- Provides an alternative to other biology frameworks
  - similar concepts to BioJava or BioPerl
  - takes advantage of Microsoft developer tools and .NET
  - will evolve as Microsoft and other contributors add features
- Designed to manipulate large data sets
  - in-memory compression of sequence data
  - data virtualization for sequences larger than memory
  - scalable algorithms that take advantage of multiple cores

MBF Design Goals

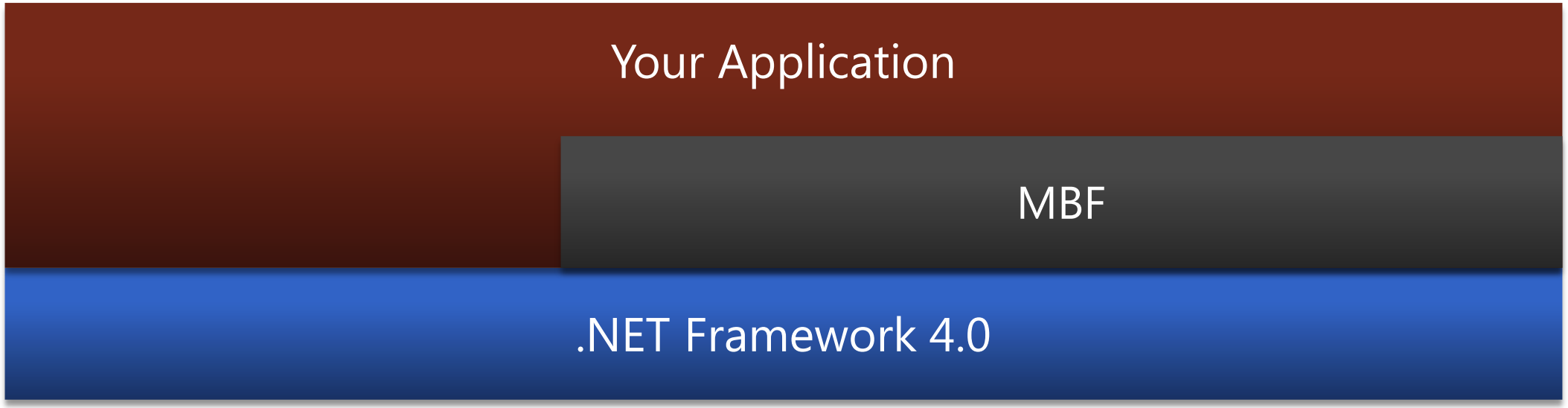
- Extensibility a primary goal
  - core concepts mapped as interfaces and ABCs
  - can easily provide alternative implementations or add any missing features you need
- Language Neutral
  - built on top of .NET – use any supported language
  - supports dynamic languages such as Python
- Designed and implemented using best practices
  - commented source code provided so nothing is a black box
  - algorithms all cite publications
- Interoperability
  - code can be run on several mainstream platforms

Architecture: Namespaces

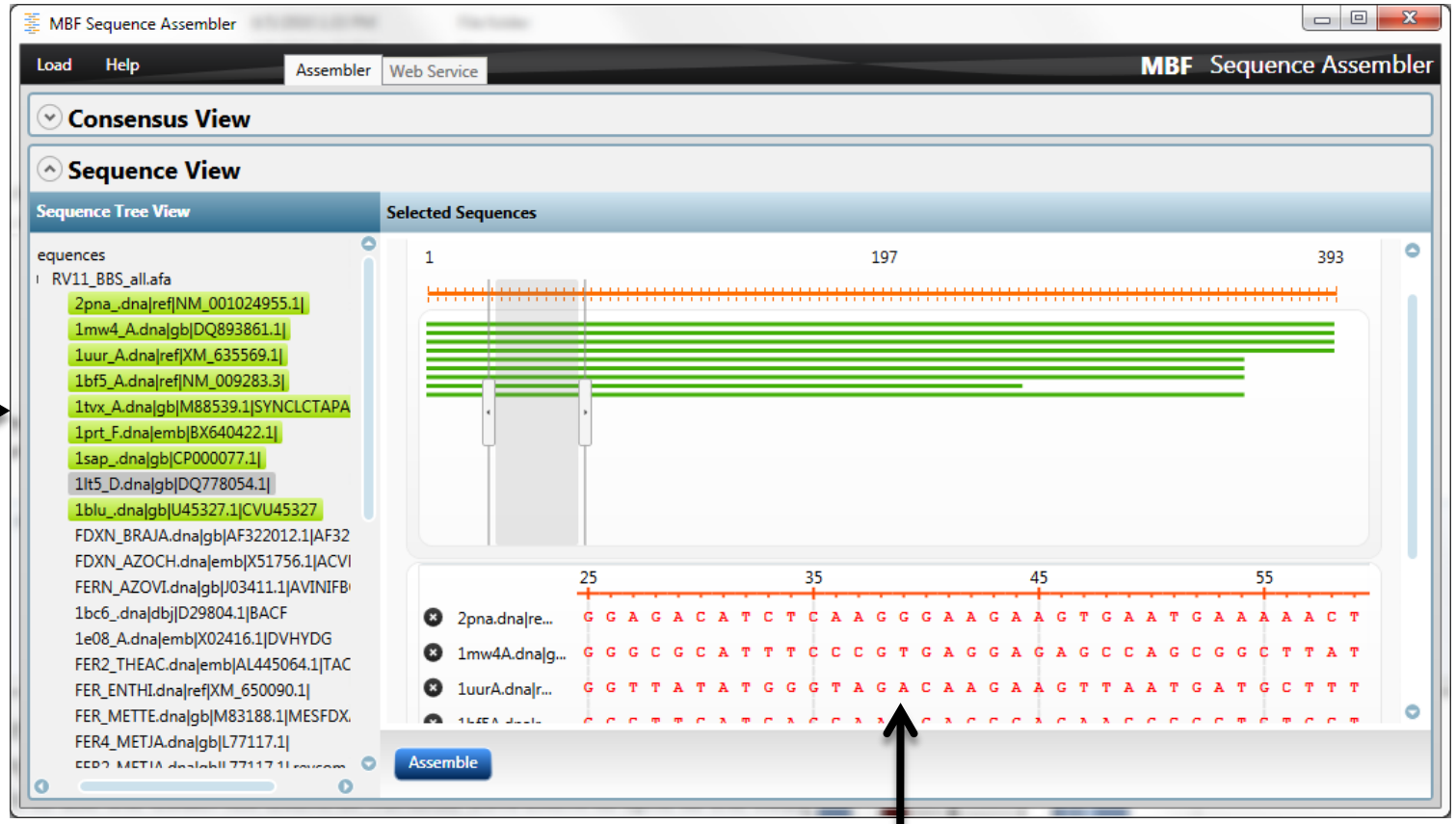


MBF vs. your application

- MBF is not an application in itself
  - it does not provide any *visualization* of the data being managed
  - it provides the *basis* for visualizations to be built on top of



Example: Sequence Assembler

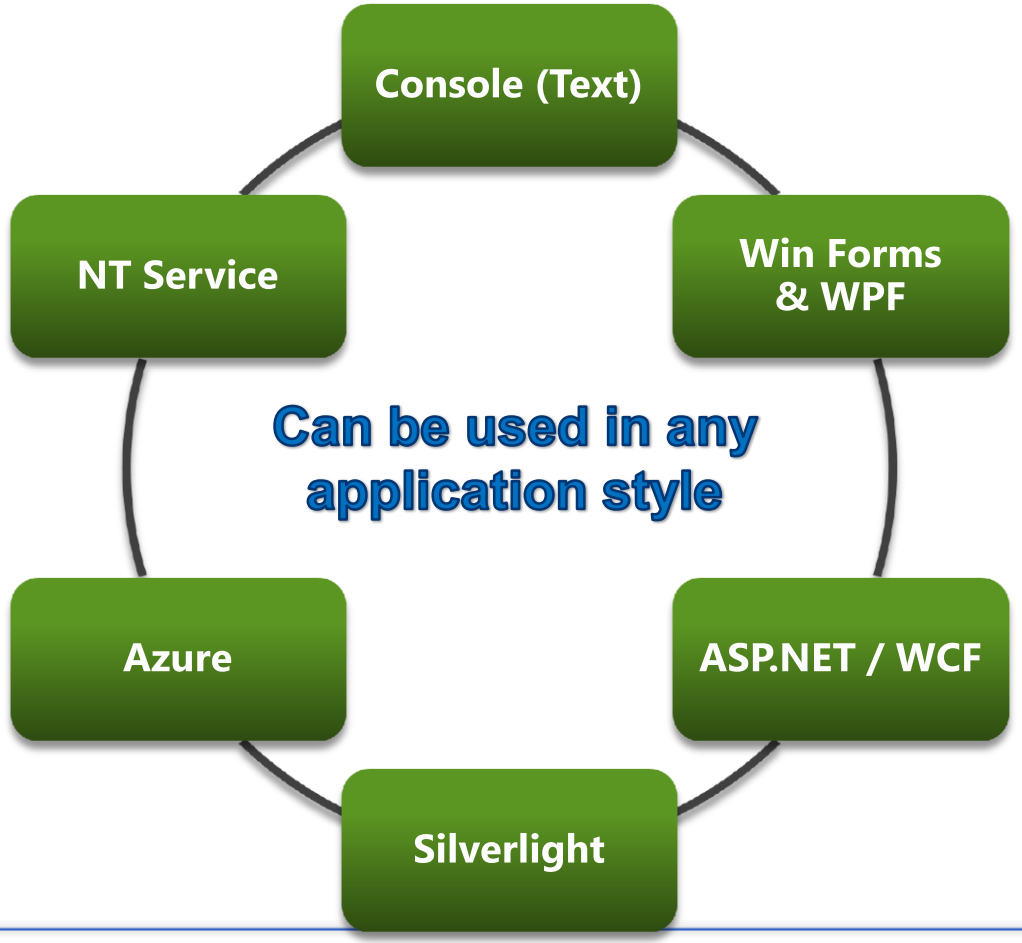


sequence data is loaded from FASTA file and assembled using MBF

drawn as nucleotide symbols and graphics using WPF

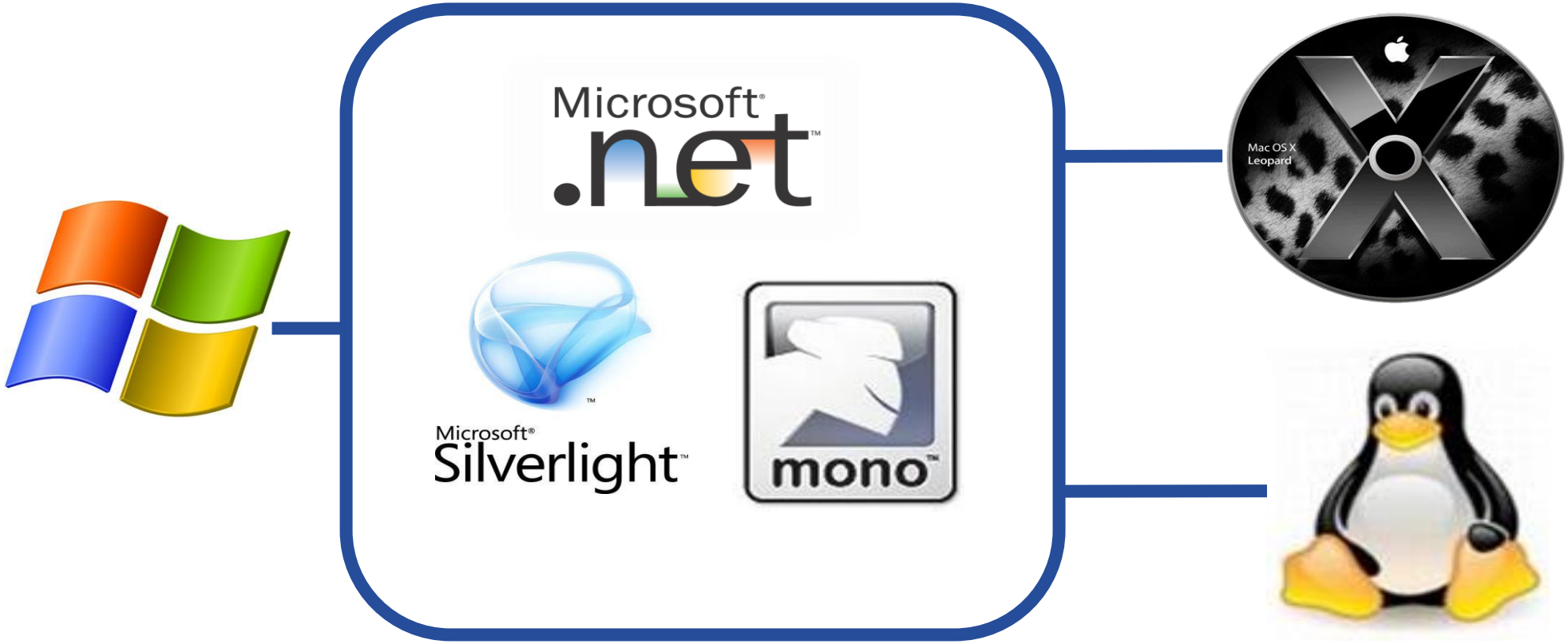
Creating Applications with MBF

- MBF allows you to work with your data however you need



Deploying your applications

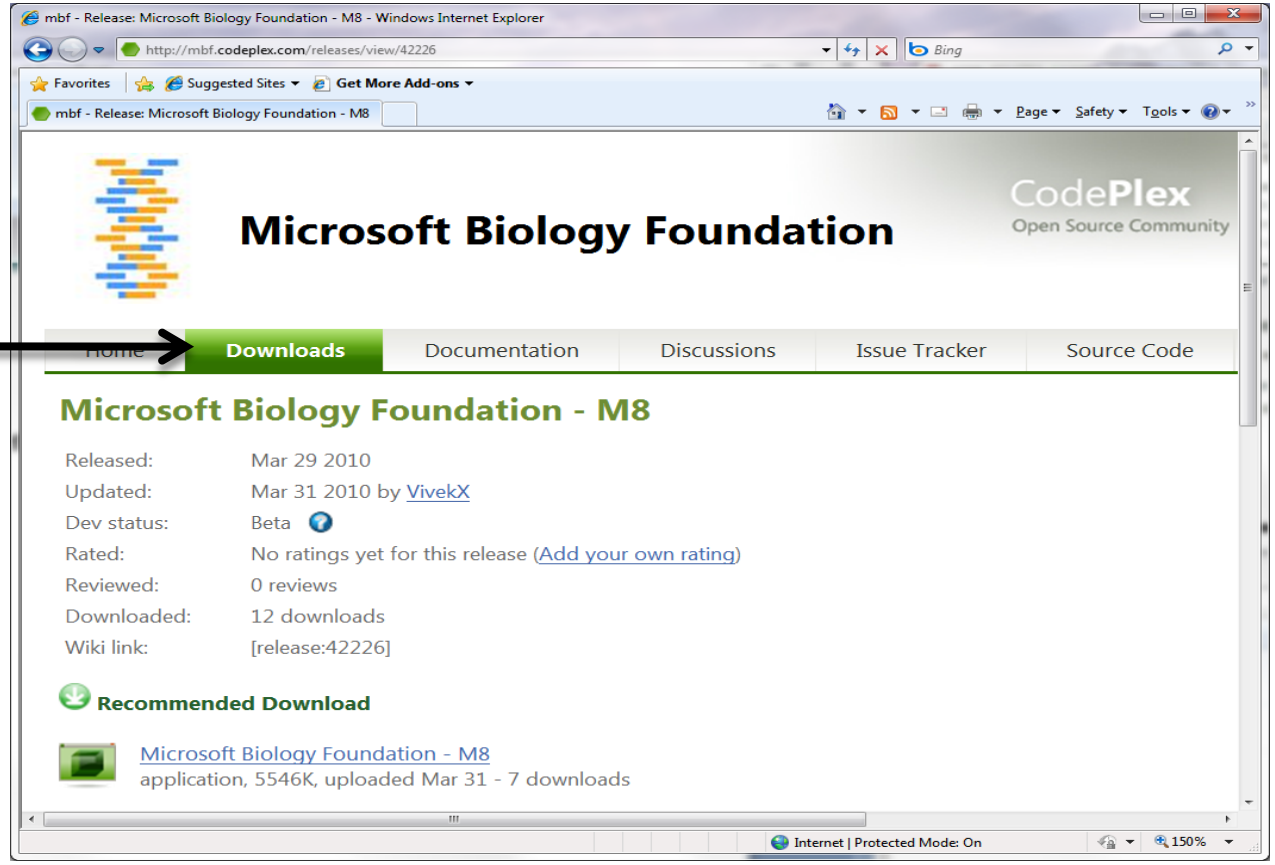
- Possible to target non-Windows platforms[1]
  - Using Silverlight / Mono / Moonlight



Getting MBF

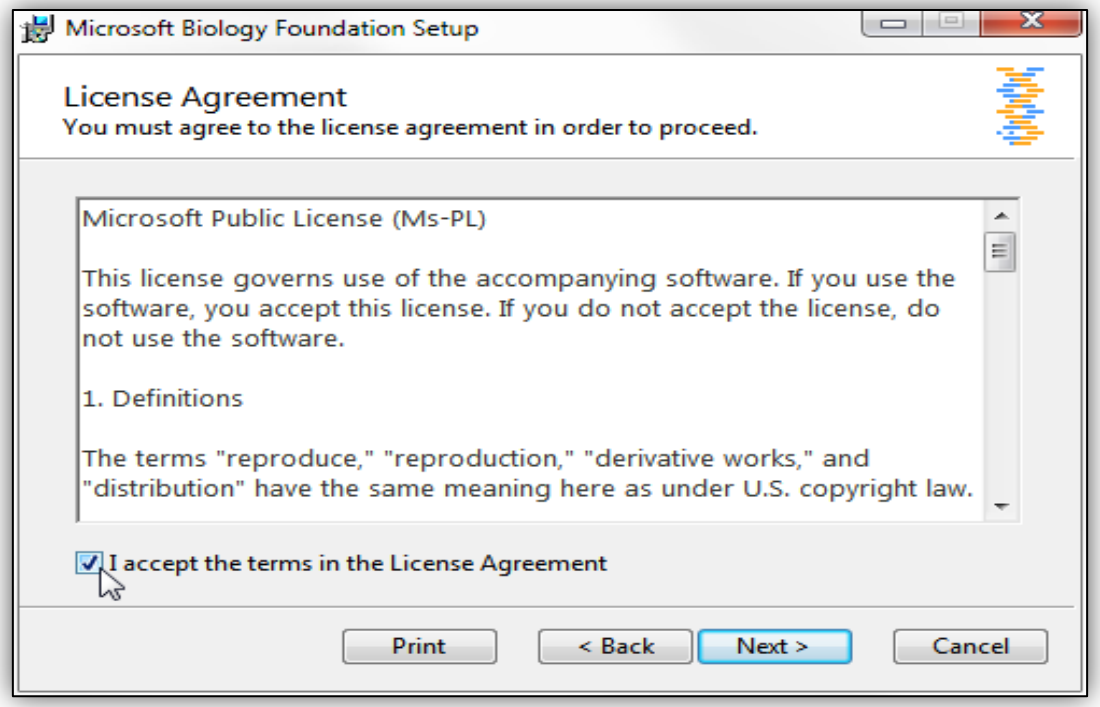
- MBF is available as an open source, free download
  - http://research.microsoft.com/bio

Downloads section lists most recent build



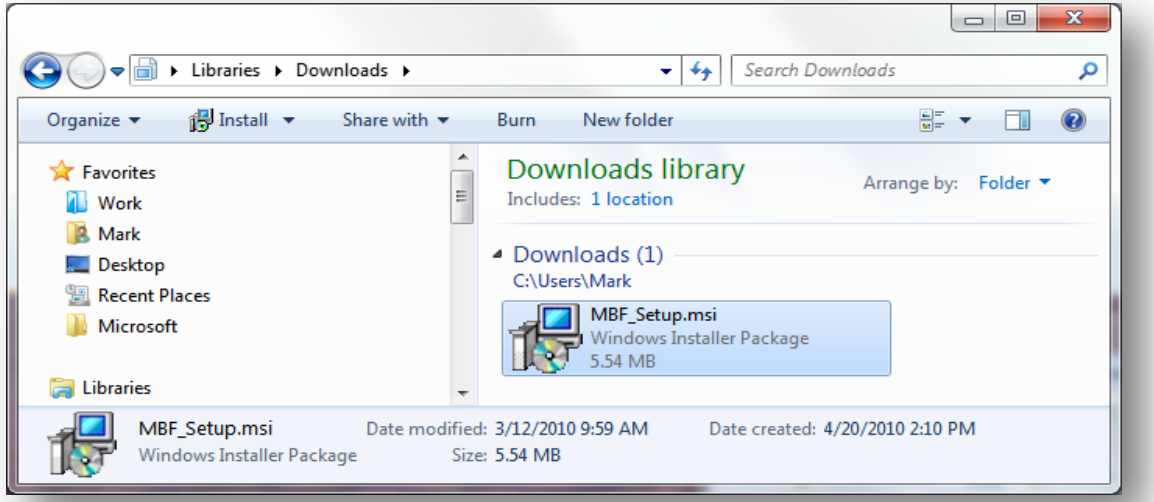
MBF Licensing

- MBF is licensed under Ms-PL
  - http://msdn.microsoft.com/en-us/library/cc707818.aspx
  - Allows you to take the code and use it in academic or commercial products



Installing MBF

- Official releases are packaged as Setup Files
  - include all the pre-built assemblies you can use immediately
  - installs full .NET 4.0 framework if not already installed
- Several other tools available
  - Sequence Assembler sample
  - Excel add-in (http://bioexcel.codeplex.com/)



Documentation

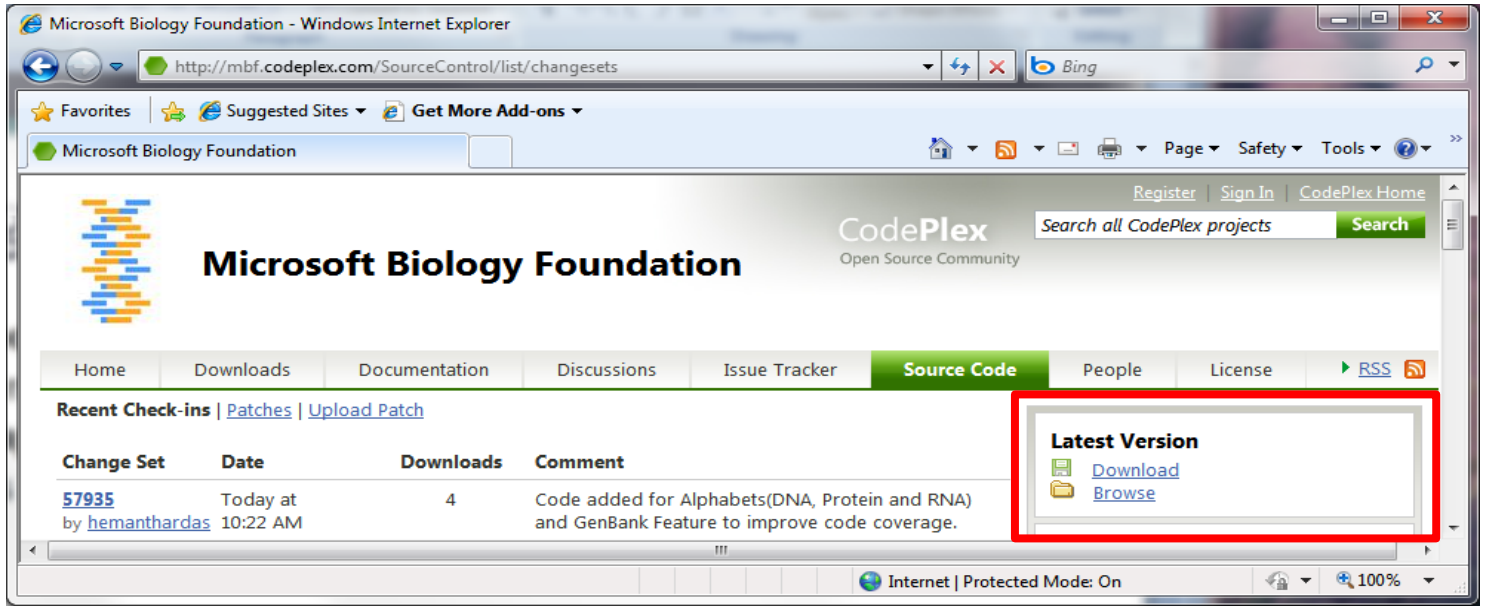
- Several documents supplied with installation (in `/Doc`)
  - even more available from http://mbf.codeplex.com/documentation
- Two documents are required reading before you begin
  - start with the **`MBF_Overview.docx`**
  - then read the **`Programming_Guide.docx`**
- `BioDotNet.chm` help file provides API reference
  - installed with SDK (full install)

`MBF_Overview.docx`
`MBF_Programming_Guide.docx`

`BioDotNet.chm`

Download the source code

- Source code is also available online at CodePlex
  - can download as a pre-packaged .ZIP file
- Can also apply to contribute to the framework
  - provides TFS credentials to get access to repository



Open Source Strategy – not always straightforward

- CodePlex.com
  - Single-branch development
  - Fully embraced in v1 project
  - But lacking in many areas
- CodePlex Foundation
  - Not planned for v1
  - Evaluating for v2
  - Working with Foundation, MS and TAB to make final decision
- Licensing
  - MBF: MS-PL
  - Sequence Assembler: MS-PL
  - MSR Biology Extension for Excel (BioExcel): MS-LPL
  - ShoRuntime: Custom

Azure Strategy

- Consumption of web services
  - AzureBLAST – prototype available from Microsoft Extreme Computing group
    - Showcasing Client + Cloud with MBF, Excel and Azure integration
  - BioHPC: Cornell's offering of bio web services on HPC
    - Cloud-like services, likely to migrate to Azure
- Looking forward
  - Significant interest from customers and partners in developing bioinformatics cloud-based services – *we would like to know more*
  - Cloud based services likely key part of follow-on project

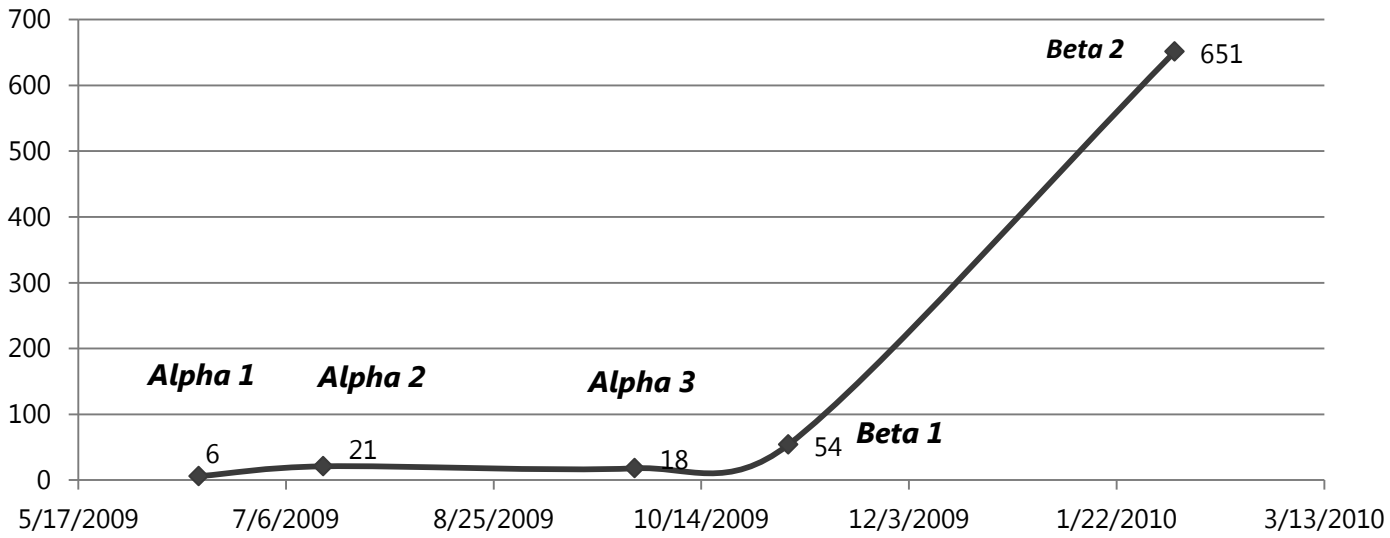
Community Activity

- **Technology Advisory Board**
  - Jarek Pillardy, Cornell
  - Kishore Doshi, UTA
  - Jim Hogan, QUT
  - Kirt Haden, Illumina
  - Jeremy Kolpak, J&J PRD
  - Vivek Kumar, Aditi
  - *Simon Mercer, Microsoft*
  - *Bob Davidson, Microsoft*
  - *Michael Zyskowski, Microsoft*
  - **Established 5/17/10**

- **Adopted by…**
  - Johnson and Johnson Pharmaceutical Research
  - Illumina Corporation
  - University of Washington
  - MSR eScience Group
- **Contributions from…**
  - Cornell BioHPC
  - Johnson & Johnson PRD
  - Queensland University of Technology
  - Illumina Corporation
  - Microsoft enthusiasts

Community Uptake and Sentiment

- Downloads to Date
  - 3800
- External Feedback
  - 23 Bugs Filed
  - 30 Discussion Posts
  - Algorithm Quality and Performance Analysis
  - 6 External Contributions

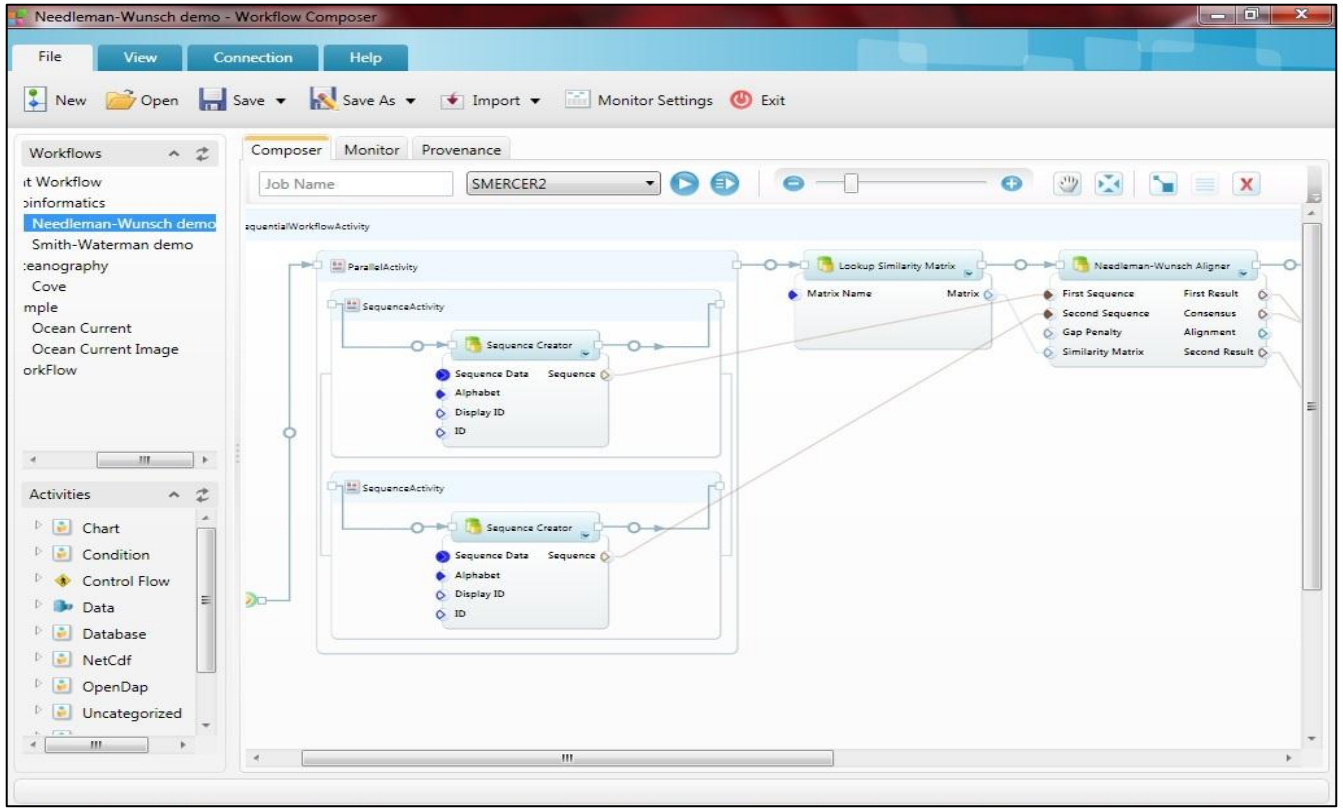


- Quotes

- "Beautiful Venn diagram :-) Thanks again, access to the developers is nice :)" Can Alkan, University of Washington
- "*MBF saved us months of development time.*" Jeremy Kolpak, J&J PRD
- "*There is clearly the opportunity to re-use much of the work you did with MBF*" - Christian Geuer-Pollmann, VENUS-C
- "*For us this is a wonderful opportunity to have a richer platform for our scientific work*" - Scott Kahn, CIO, Illumina Inc.
- "*The timing is perfect for this – if you had tried it three years ago it wouldn't have worked*"
  - Rob Armstrong, VP Global External R&D, Eli Lilly
- "*This is really good work*" – Nuzrul Haque, Pfizer

Better Together: Trident Scientific Workflow Workbench

A visual workflow environment that allows researchers to better manage, evaluate and interact with even the most complex scientific datasets



- Built on top of Windows Workflow Foundation
- Write once, deploy and run anywhere…
- Visually program workflows
- Libraries of activities and workflows
- Automatic provenance capture

Available at: http://tridentworkflow.codeplex.com/

mbf.codeplex.com

- Open Source
  Available free of charge for commercial and non-commercial use and modification under the MS-PL license (http://opensource.org/licenses/ms-pl.html)
- Community-Developed
  Moved to CodePlex, Created advisory board and building community
- Community-Curated
  Modify code, find bugs, contribute new features
- Periodic Releases
  Snapshots of open source with additional testing

Microsoft® Research
Faculty Summit 2010

demo

Demos

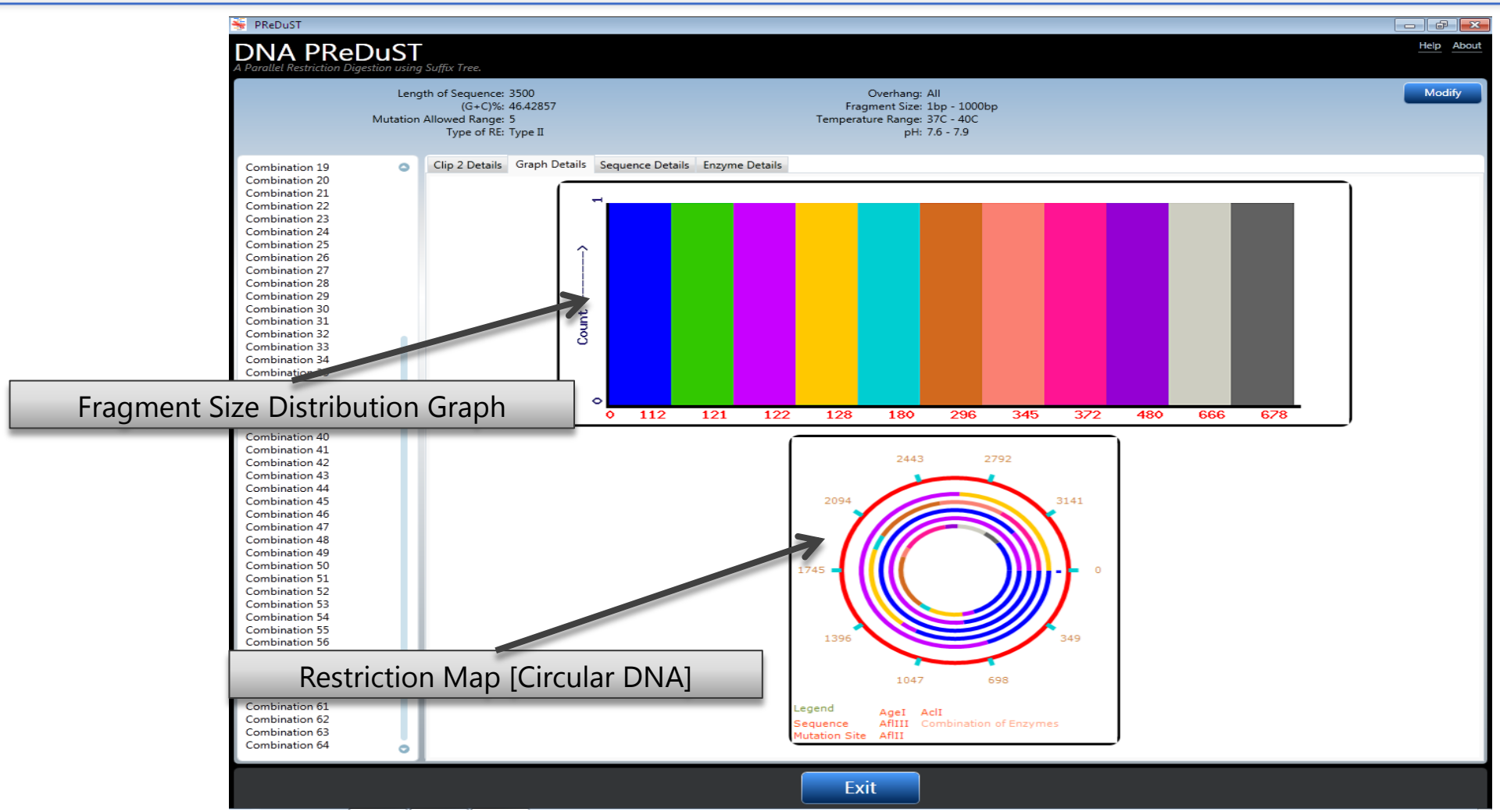

TestAddIn.xlsx - Microsoft Excel (Technical Preview)
Home
Insert
Page Layout
Formulas
Data
Review
View
Bio Informatics
Load Test
Team
Import From
Export To
Sequence Data
Select BLAST service
Search Status:
Done
Cancel Search
Web service
3D Molecular Viewer
Visualization
Select Aligners
Alignment Status:
Done
Cancel Alignment
Sequence Alignment
Charts
Charting
Configure
Configuration
Microsoft Biology Foundation
GW2
R
Seq_1
Consensus
Ready
MBF Sequence Assembler
File
Options
Help
MBF Sequence Assembler
Sequences
Consensus
Consensus Tree View
Contigs
Contig1
[Sequence1 Read 26]
[Sequence2 Read 21]
[Sequence3 Read 10]
[Sequence4 Read 8]
[Sequence5 Read 11]
[Sequence6 Read 23]
[Sequence7 Read 20]
[Sequence8 Read 2]
[Sequence9 Read 19]
[Sequence10 Read 6]
[Sequence11 Read 1]
[Sequence12 Read 7]
[Sequence13 Read 17]
[Sequence14 Read 18]
[Sequence15 Read 22]
[Sequence16 Read 16]
[Sequence17 Read 3]
[Sequence18 Read 4]
[Sequence19 Read 13]
[Sequence20 Read 12]
[Sequence21 Read 9]
[Sequence22 Read 25]
[Sequence23 Read 14]
[Sequence24 Read 24]
[Sequence25 Read 5]
[Sequence26 Read 15]
Report
Custom View
Consensus
Sequences
Reverse Complemented
Complimented
Reversed
1
905
1809
Executing Search...
This could take a while. Please wait.
Cancel
Blast

Microsoft® Research

Faculty Summit 2010

partners

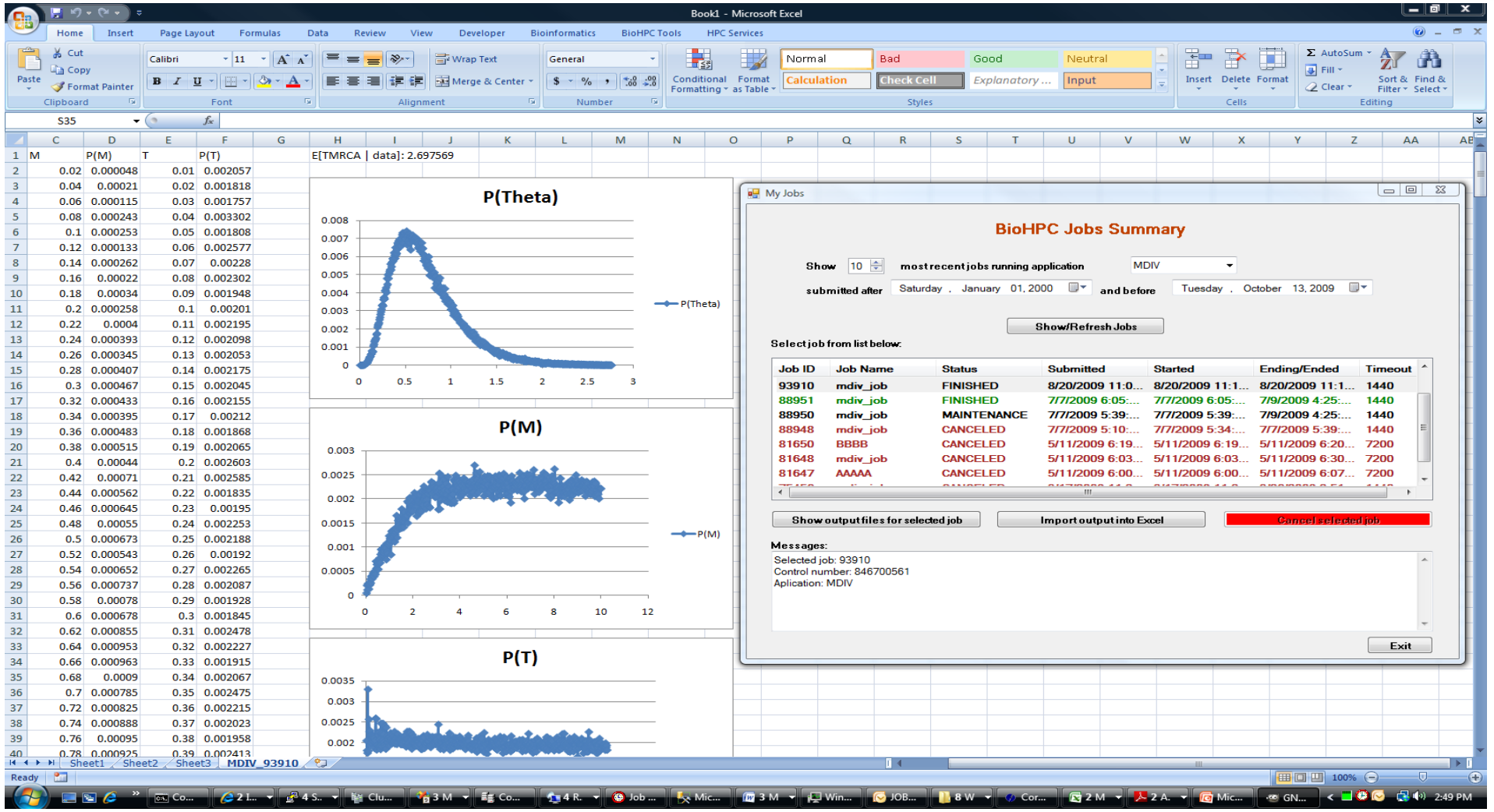
Selecting Restriction Endonucleases: DNA PReDuST



Aditi Technologies

Computational Biology Applications Suite for High Performance Computing (BioHPC)

Computational Biology Service Unit



Cornell University

Acknowledgements

- **MBF Team**
  - Mike Zyskowski, Chris Wu
- **Microsoft Research**
  - David Heckerman, Bob Davidson, Carl Kadie, Yogesh Simmhan, Jennifer Listgarten, Jonathan Carlson
- **Cornell University**
  - Jarek Pillardy
- **Queensland University of Technology**
  - Jim Hogan
- **University of Texas at Austin**
  - Robin Gutell
- **Aditi Technologies**
  - Vivek Kumar
- **Illumina Corporation**
  - Scott Kahn
- **Johnson & Johnson Pharmaceutical Research Division LLC.**
  - Dimitris Agrafiotis, Victor Lobanov, Jeremy Kolpak

http://research.microsoft.com/bio/

Microsoft® Research

Faculty Summit 2010