

DRAFT (Feb 17, 2004)
caBIG Workspace Developer Project Form

Developers, please complete this form in advance of the caBIG kickoff meeting and return by e-mail to adamsm@mail.nih.gov. Completed forms will be made available to all participants in advance of the meeting to enhance workspace discussions. During our conversations with you, we expressed the aspect of your program that we would like you to develop in the first year of the caBIG pilot; it is this we are asking you to address - here and in your presentation.

1. Sponsoring Cancer Center

Lombardi Comprehensive Cancer Center at Georgetown University

2. Workspace

Integrative Cancer Research Workspace

3. Project or Activity

Tool development for comprehensive analysis of very high dimensional data spaces
Tool deployment and adaptation, e.g., tools in our Visual Statistical Data Analyzer (VISDA).

4. Workspace needs the project meets

Tools for microarray data preprocessing, for the visualization and analysis of very high dimensional data, computational correction of tissue heterogeneity, biomarker identification, and for phenotype classification and prediction.

5. Stage of project maturity (Conceptual, early beta, regular end-user use at parent center, regular use in the community)

Tools at various stages of development. Most of the components of VISDA are already in use at parent center and other sites. Newer components and updated components are in beta testing/optimization.

6. Technical details of Tools

- a. Software Architecture (These will likely be preliminary)
 - i. System design
 - ii. Component details
 - iii. Relevant standards
 - iv. UML schematics (if valid)
 - v. Size of project installed software base
- b. Development Environment (tools, languages, bug tracking, etc.)

6. Technical details of Tools

We have been actively developing a set of computational tools for microarray data analysis for cancer research, with a joint effort from Georgetown University, Catholic University of America and Virginia Tech. The tools can be functionally organized into three categories:

- 1) *Data Preprocessing* – Cross-Phenotype Normalization, and Tissue Heterogeneity Correction;
- 2) *Classification & Prediction* – Optimized Multilayer Perceptrons (MLP) classifiers, and Adaptive Hierarchical Subspace Experts (AHSE);
- 3) *Cluster Discovery & Visualization* - Visual Data Analysis (VISDA) Package.

For Data Preprocessing and Classification, we have developed a comprehensive *Multitask Gene Selection* tool to help (1) alleviate the “curse-of-dimensionality” problem in classification and prediction, and (2) identify different gene sets for different tasks (e.g., pathway building, normalization, tissue heterogeneity correction, classification, etc.) in microarray data analysis.

a. Software Architecture

i. System design

Fig. 1. illustrates a block diagram of our system design, which integrates a variety of Matlab toolboxes with C++ software packages (dChip and OpenInventor), and the R package.

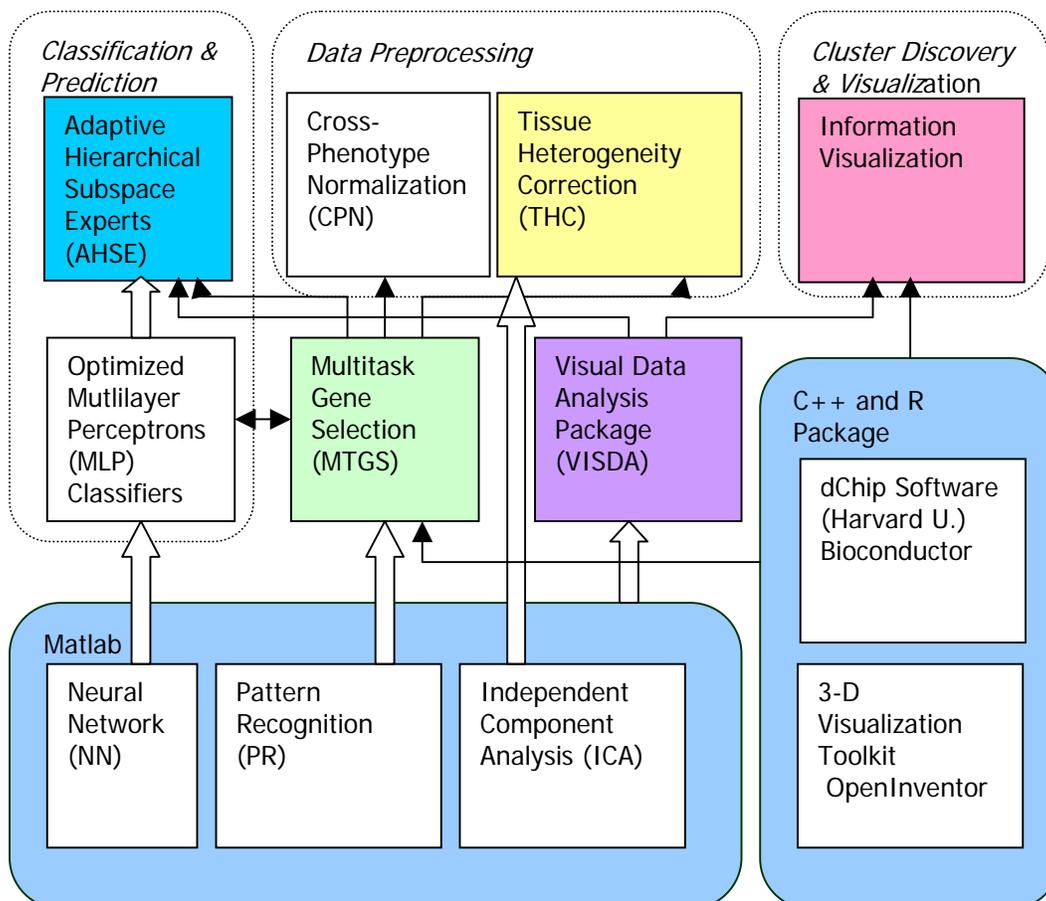


Fig. 1. Block diagram of our system design.

ii. Component details

Three major components are detailed below to illustrate their modularity. They are (1) VISDA component, (2) MTGS component, and (3) AHSE component.

i. **VISDA COMPONENT**

The VISDA software has ten sub-modules that can be categorized into three functional groups: (1) Top-Level Group, (2) Sub-Level Group and (3) Visualization Group. An overview of the VISDA modules is given Fig. 2, where the relationship of those three functional groups is also flow-charted. Top-Level Group has two modules: (1) VE_INIT and (2) VE_MSELECT. Sub-Level Group includes three modules: (1) VE_SUB_NEW, (2) VE_SUB_EM and (3) VE_MSELECT2. Visualization Group has five modules: (1) VE_MVIS, (2) VE_VIS_T, (3) VE_VIS_CIR, (4) VE_FSIZE, and (5) VE_TICK.

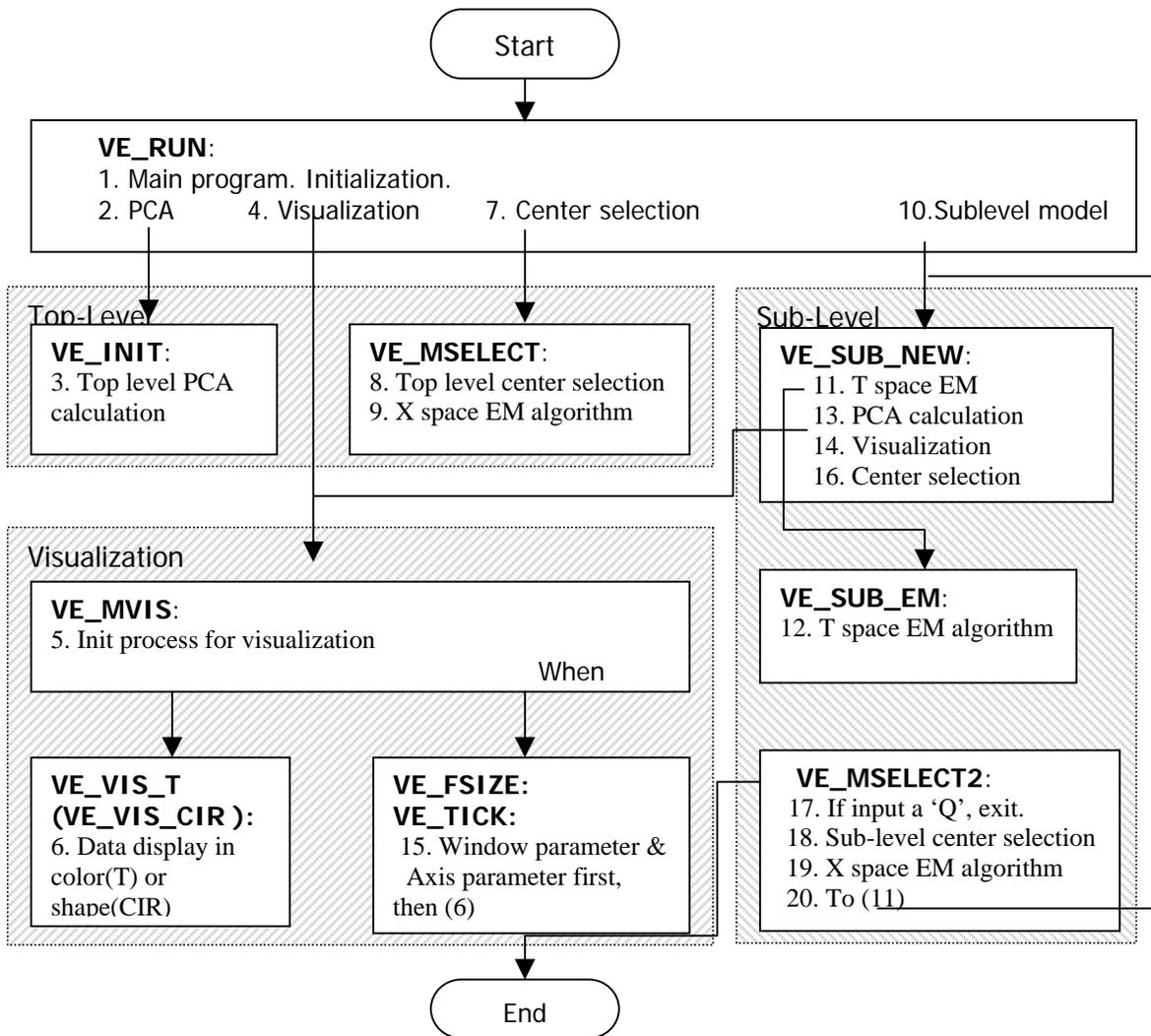


Fig. 2. An overview of the VISDA component.

ii. MTGS COMPONENT

The MTGS component consists of three modules: (1) Constantly-Expressed Gene Selection, (2) Independence Support Gene Selection, and (3) Discriminatory Gene Selection (DGS) for Classification and Prediction. Fig. 3 illustrates the DGS module that including the following submodules: (1) individually discriminatory gene (IDG) selection by a weighted Fisher Criterion (wFC), (2) jointly discriminatory gene (JDG) selection by a wFC and a sequential floating search method, (3) leave-one-out procedure, and (4) neural network classifiers.

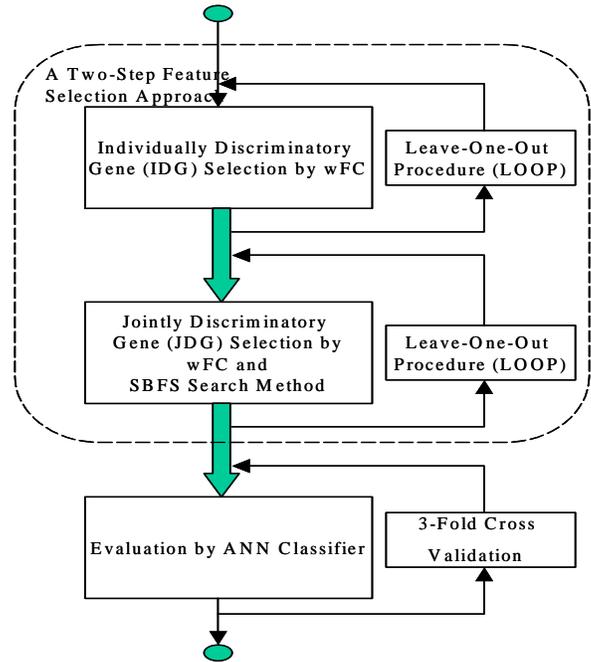


Fig. 3. Block diagram of the Discriminatory Gene Selection Module (DGS).

iii. AHSE Component

The AHES component consists of two major modules: (1) tree-construction by VISDA, and (2) MLP subspace experts. An example of tree-structured MLP subspace experts is illustrated in Fig. 4. For multiclass prediction problem, AHSE has been demonstrated to be able to offer us an improved performance over conventional MLP classifiers.

b. Development Environment (tools, languages, bug tracking, etc)

The primary purpose of our development was to support our research projects. As shown in our system design (Fig. 1), majority of the tools are developed based on Matlab environment. Recently, we have started to migrate our algorithm development to C/C++ environment. Particularly, the normalization and gene selection algorithms have already been integrated into dChip software using Visual C++ and R languages. We believe that caBIG compatible APIs can help migrate all our tools into caBIG systems or environment.

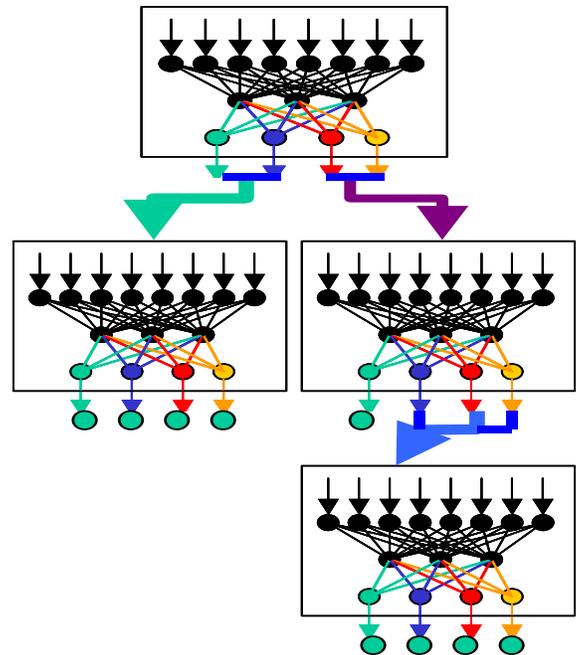


Fig. 4. Block diagram of the AHSE component.

Bioimaging & Bioinformatics for Integrative Cancer Research

Also, we are actively developing an integrative approach for breast cancer research – Integrating *in vivo* functional molecular imaging with *in vitro* and *in vivo* molecular analyses. An overview of our integrative approach is illustrated in Fig. 5. Our group has developed several image analysis toolboxes – (1) Model-Based Image Segmentation, (2) Deformable Image Registration, 3-D Modeling of Tissue Components in Breast, and (4) Independent Component Analysis for Blood Flow Characterization. The molecular analysis system is interplayed with imaging toolboxes for this research component (as illustrated in Fig. 1). We are currently focusing on integrating all the toolboxes into a comprehensive research tool – Functional/Molecular Imaging of Composite Signatures of Breast Cancer.

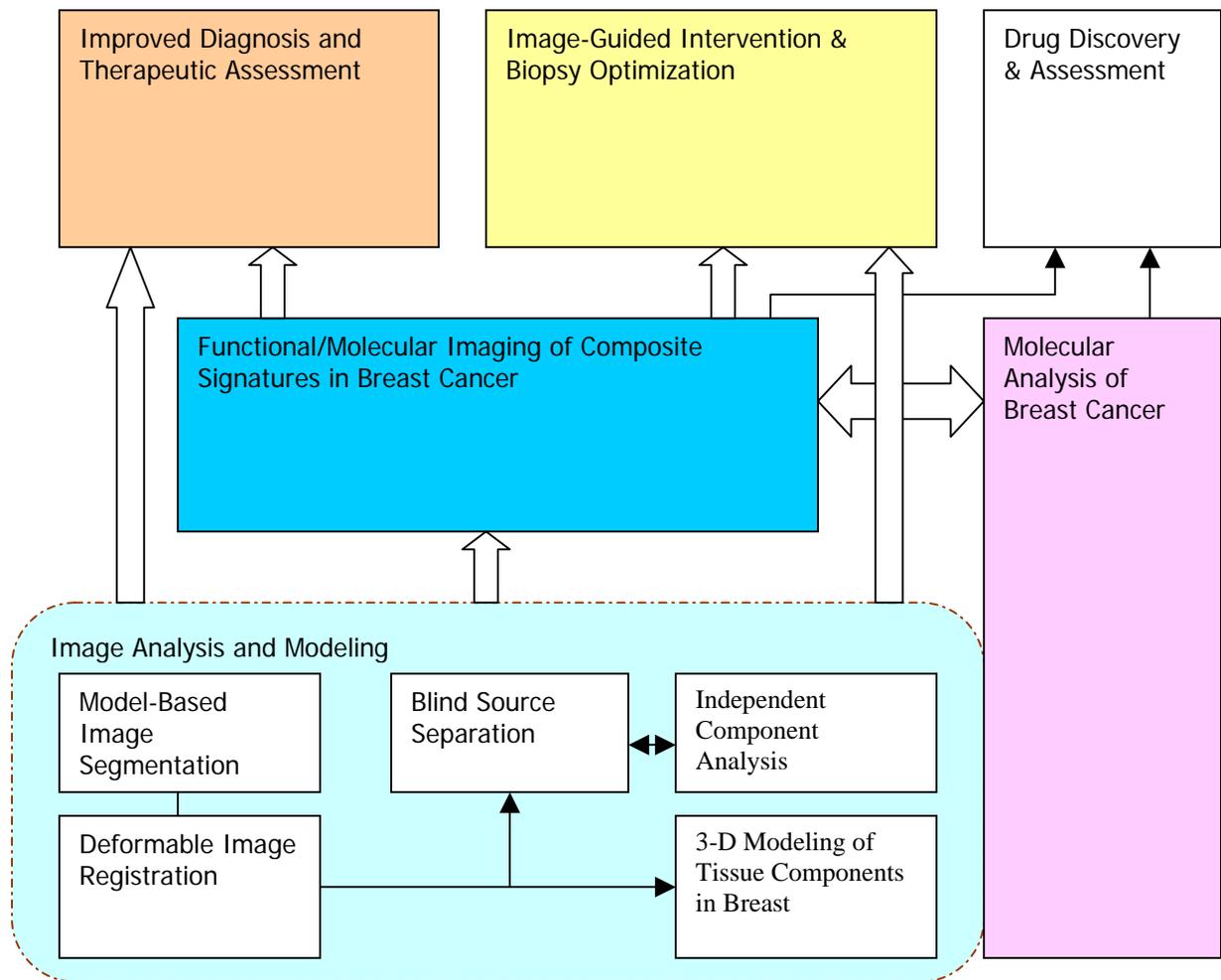


Fig. 5. An overview of the integrative approach – Integrating *in vivo* molecular imaging with *in vitro* molecular analysis.

7. Does the project make use of existing standards? If so, what are they?
(e.g. bioinformatics standards such as MIAME for microarrays, or software standards such as XML)

Yes but some additional modifications may be necessary to integrate with standards adopted by caBIG participants

8. Does other software in the community meet this need? Is this software open source? Can it be harnessed?

No other software implements the algorithms we have developed. All is open source code and can be harnessed

9. Points of possible interoperability with other caBIG systems
(This might include communication with other caBIG databases, use of caCORE APIs, caBIG-compatible APIs, etc.)

Once standard databases and APIs are adopted, engineering full interoperability should be straightforward.

10. What resources are proposed to achieve caBIG interoperability?

- a. Developmental requirements
 - i. Software (re)engineering
 - ii. Standards adoption
 - iii. Platform migration

Funds for supporting professional programmers working on translation of algorithms into reliable and portable software package, software documentation, quality control and assurance.

- b. Infrastructure
 - i. Facilities
 - ii. Management tools
 - iii. Personnel

Funds for C++/Java development environment, (UNIX/Linux server, high-end PCs, software system administrator).

11. Draft 12-month work plan, with milestones to achieve caBIG interoperability.

- 1-3 months VISDA caBIG interoperability
- 4-6 months CPN and PICA-THC caBIG interoperability
- 7-9 months MTGS and MLP/AHSE caBIG interoperability
- 10-12 months Software system integration and final testing