

# User's Guide to Product Comparison Spreadsheets

## Introduction

Product Comparison Spreadsheets are useful tools for comparing data generated from two expression arrays. These spreadsheets can be used to find relationships between probe sets for similar sequences from two different products. We currently offer two types of Product Comparison Spreadsheets:

- 1) Product Family Comparison
- 2) Cross-Species Comparison

The product family comparison looks at two designs from the same product family. Due to the dynamic nature of the public databases, probe sets between different versions of a product family, such as the Human Genome arrays, will not be identical. In some cases the same sequences will be represented by completely different probe sets, creating a challenge when comparing data sets generated on different generations of a product family. These spreadsheets allow some level of data comparison as the product line evolves.

The cross-species comparison looks at two designs from different product families, such as the Rat Genome arrays versus the Mouse Genome arrays. The comparison spreadsheets can be used to find highly similar sequences between the two species.

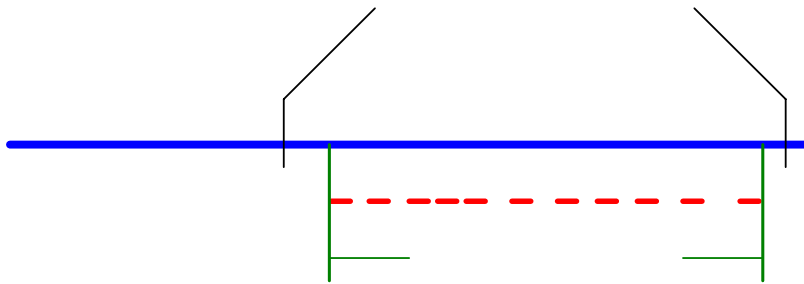
This document describes the four different files that Affymetrix provides for both Product Family and Cross-Species comparisons. Each file offers a different level of comparative analysis.

## An Overview of the Probe Selection Process

To understand how selection criteria are applied to generate the spreadsheets, a brief description of our probe selection process and terms is required. The probe selection process begins with the identification of a representative sequence for a UniGene cluster or Affymetrix subcluster. Depending on the array, the representative sequence is either a consensus or exemplar sequence.

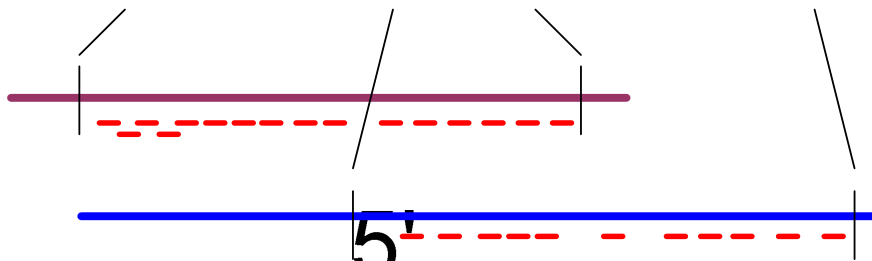
Next, a probe selection region is defined, which is normally the last 600 bases from the 3' end. The probe selection algorithms are then used to select probes. The region of sequence that encompasses the most 5' to the most 3' probes is labeled as the Target Sequence. Since these sequences are stored in a Sequence Information File (.SIF file), they are also referred to as the SIF Sequences.

Figure 1 shows a consensus sequence as the representative sequence. The probe selection region measures 600 bases from the 3' end. The Target Sequence (SIF) is a sub-region within the Probe Selection Region.



**Figure 1**

Figure 2 illustrates an example of how sequences are matched between two array designs. Sequence A from the first design is aligned with Sequence B from the second design, with the corresponding probe selection regions displayed for both.



**Figure 2**

Based on this example, the following values would be returned to the comparison spreadsheet:

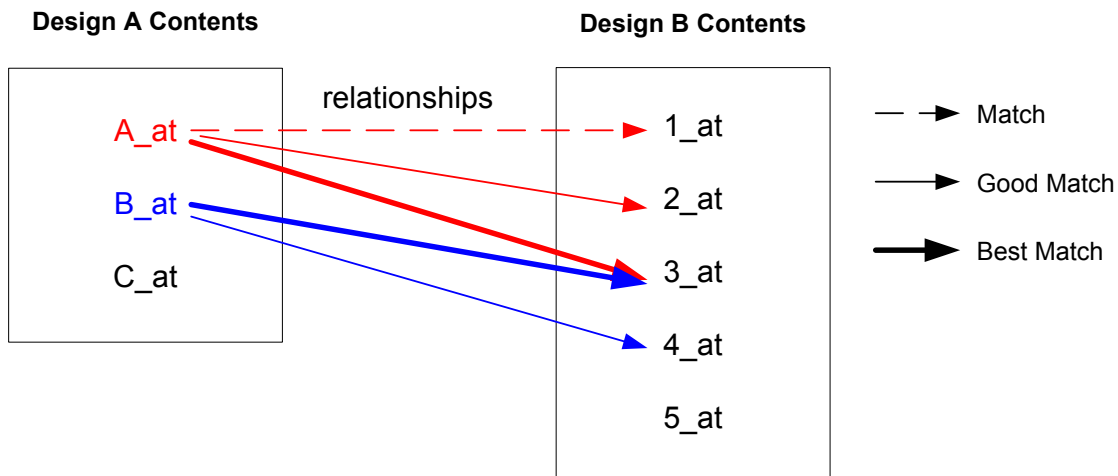
Column Heading	Value Returned	Definition
A Matches B Probe Selection Region	6	The number of Perfect Match probes from Sequence A which align to the probe selection region of Sequence B
B Matches A Probe Selection Region	5	The number of Perfect Match probes from Sequence B which align to the probe selection region of Sequence A
A Matches B Representative	16	The number of Perfect Match probes from Sequence A which align to any region of Sequence B
B Matches A Representative	6	The number of Perfect Match probes from Sequence B which align to any region of Sequence A

## An Overview of the Comparison Spreadsheets

There are up to four different tab-delimited text files available for a given design: Complex, Good Match, Best Match, and No Match. The latter three files are generated from the Complex spreadsheet, which is discussed in greater detail later in this document.

**PLEASE NOTE:** The Complex spreadsheet is provided for advanced users who want to design their own queries and fully understand the special considerations necessary to deal with potentially very large text files (for example, the human array comparison data is ~200 MB unzipped, containing 2.2 million rows).

For the Complex, Good Match, and Best Match spreadsheets, there are twenty different column headings. “A” and “B” in the column headings refer to different designs. For Product Family Comparison, the first design “A” always refers to the previous design of the same product family and “B” refers to the current design. Design “A” acts as the reference point for all comparisons. For example, a comparison between the two human designs HG-U95 (design A) and HG-U133 (design B), looks at each probe set in HG-U95 and finds the corresponding matches in HG-U133, but not the other way around. Here is a graphical representation of what the different files contain:



The following entries will appear in the different spreadsheets

Complex		Good Match		Best Match		No Match
A_at	1-at					
A_at	2_at	A_at	2_at			
A_at	3_at	A_at	3_at	A_at	3_at	
B_at	3_at	B_at	3_at	B_at	3_at	
B_at	4_at	B_at	4_at			C_at

A description of the different column headings is presented in the table below:

### **Comparison Spreadsheet Column Headings**

<b>Column Heading</b>	<b>Description</b>
A Array Name	The name of the first array design.
A UniGene Cluster	The UniGene cluster from which the representative sequence for a probe set from the first design was derived.
A Probe Set Name	Probe set name from the first design.
B Array Name	The name of the second array design.
B UniGene Cluster	The UniGene cluster from which the representative sequence for a probe set from the second design was derived.
B Probe Set Name	Probe set name from the second design.
Percent Identity	Percent identity when the representative sequences from Probe Set A and Probe Set B are aligned.
Match	The number of matching bases when the representative sequences from Probe Set A and Probe Set B are aligned.
MinLen(Rep(A),Rep(B))	The length of the shorter representative sequence between the two designs. This value indicates the maximum possible overlap between the two representative sequences.
A Matches B Probe Selection Region	The number of Perfect Match probes from the first design which align to the probe selection region of a sequence in the second design.
A Total Probes in Probe Set	The total number of Perfect Match probes in Probe Set A.
B Matches A Probe Selection Region	The number of Perfect Match probes from the second design which align to the probe selection region of a sequence in the first design.
B Total Probes in Probe Set	The total number of Perfect Match probes in Probe Set B.
A Matches B Representative	The number of Perfect Match probes from the first design which align to the representative sequence for a probe set in the second design.
B Matches A Representative	The number of Perfect Match probes from the second design which align to the representative sequence for a probe set in the first design.
Target Seq (SIF) Percent Identity	The percent identity when the Target Sequences from Probe Set A and Probe Set B are aligned.
Target Seq (SIF) Match	The number of matching bases when the Target Sequences from Probe Set A and Probe Set B are aligned
MinLen(Target Seq(A),Target Seq(B))	The length of the shorter Target Sequence between the two designs. This value indicates the maximum possible overlap between the two Target Sequences.
Good Match	“Y” if the entry meets the criteria for the “Good” match table (see criteria listed in the next section).
Best Match	“Y” if the entry meets the criteria for the “Best” match table (see criteria listed in the next section).

## **Generation of the Complex Spreadsheet**

The Complex Spreadsheet was generated from the Affymetrix<sup>®</sup> array design database. It contains a list of ALL probe set pairs between the two designs that meet ANY of the following criteria:

- **Percent Identity or Sequence Overlap > 50%, where Sequence Overlap = Match / MinLen(Rep(A),Rep(B)) \* 100**
- **The value returned in A Matches B Probe Selection Region or B Matches A Probe Selection Region is > 0**
- **Target Seq (SIF) Percent Identity or Target Seq (SIF) Overlap > 50%, where Target Seq (SIF) Overlap = Target Seq (SIF) Match / MinLen(Target Seq(A),Target Seq(B)) \* 100**
- **The value returned in A Matches B Representative or B Matches A Representative is > 0**

## **Generation of the Good, Best and No Match Spreadsheets**

The Complex Spreadsheet can be very large. While it is a useful tool for GeneChip<sup>®</sup> users who want to design their own queries, many others would prefer to have predefined spreadsheets that give only the most significant probe set matches between the two designs. For those users, we have generated two additional spreadsheets from the Complex Spreadsheet: the Good Match and Best Match Spreadsheets.

### **Good Match Spreadsheet**

A probe set pair from the two designs is included in the Good Match Spreadsheet only if ALL of the following criteria are met:

- **Percent Identity > 90%**
- **MinLen(Rep(A),Rep(B)) > 100**
- **The value returned in A Matches B Probe Selection Region is > 1**
- **The value returned in B Matches A Probe Selection Region is > 1**

The Good Match Spreadsheet will produce some entries where more than one probe set from the second design is returned for each entry in the first design. In most cases, researchers will want to find the best match to a probe set in the first design. To create a list of these “best” matches, a more stringent set of parameters are applied to the Good Match Spreadsheet.

### **Best Match Spreadsheet**

The criteria shown below (in priority ranking) were applied to the Good Match Spreadsheet. In each case, the probe set with the highest returned value was added to the Best Match Spreadsheet:

- **A Matches B Probe Selection Region**
- **B Matches A Probe Selection Region**

- **Percent Identity**
- **Target Seq (SIF) Percent Identity**
- **Target Seq (SIF) Match**
- **Match**
- **MinLen(Target Seq(A), Target Seq(B))**
- **MinLen(Ref(A), Ref(B))**
- **A Matches B Representative**
- **B Matches A Representative**

For instances where more than one match was found for a probe set from the first design, an additional evaluation was performed to select only ONE probe set from the second design by using the following:

- 1) If all the tied entries were cases of different probe sets to the same sequence (i.e. gene (\_a), unique (no suffix), similar (\_s), and mixed (\_x) sets), the probe set with the highest priority was chosen. The priority is defined as: \_a, unique, \_s, \_x.
- 2) If the tied entries were probe sets to different sequences, then a manual review was performed by Affymetrix scientists to pick the best one.

### No Match Spreadsheet

Any probe sets from the first design that showed no matches to the second design in the Complex Spreadsheet were added to the No Match Spreadsheet. This spreadsheet is a simple two-column list including the array name and the probe set name from the first design. If every probe set has at least one match in the second design, this spreadsheet will not be provided.

### **Guidelines for Using the Comparison Spreadsheets**

The spreadsheets described here are available as downloads from the Affymetrix® website ([www.affymetrix.com/support/technical/comparison\\_spreadsheets.affx](http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx)) as tab-delimited text files. Once downloaded, users can easily import the Good, Best, or No Match spreadsheets into either Microsoft Excel or Microsoft Access, or leave them as text files for uses in other data analysis applications.

As stated previously, the Complex Spreadsheet is provided for advanced users who want to design their own queries and fully understand the special considerations necessary to deal with very large text files. The visualization tools at [www.affymetrix.com](http://www.affymetrix.com) are an alternative available to users who want to compare probe sets between two different designs, one set at a time.

## **Glossary**

Consensus sequence: The result of a consensus calling method from the alignment of contiguous, clustered sequences such as a UniGene cluster or subcluster created by Affymetrix.

Exemplar sequence: An individual sequence chosen to represent a UniGene cluster or subcluster created by Affymetrix.

Representative Sequence: Refers to the corresponding consensus or exemplar sequence used for probe set alignments.

3' end: Defined as an exemplar end, a consensus end, or an alternative polyadenylation site within a consensus defined by a stack of EST ends.

Probe Selection Region: Region of the transcript from which probes are selected. Often, this region corresponds to approximately the 600 most 3' bases.

SIF: A text file containing the Target Sequences for a given design.

Target Sequence: After probe selection, the region of the transcript that encompasses the most 5' and 3' probes. They are also called SIF sequences.

Perfect Match: A single stranded DNA oligonucleotide that is complementary to the target. For antisense-target detecting arrays, the Perfect Match oligo is identical to the designated forward strand of the gene sequence it represents. For sense-target detecting arrays, the Perfect Match oligo is complementary to the designated forward strand of the gene sequence it represents.

Mismatch: A single stranded DNA oligonucleotide that is identical to the Perfect Match probe, except for a single base substitution at the central position.

Probe: 25-base oligonucleotide selected by Affymetrix for synthesis on the array.