Next-generation analysis of deep sequencing data:

Bringing light into the black box of phage display experiments

Dr. Michael Blank (CSO, AptalIT)
Accelerating *in vitro* selection procedures

NGS and *in silico* analysis in the context of *in vitro* evolution procedures
The application fields of COMPAS *

1) design & synthesis of starting library
2) in vitro selection
3) Next Generation Sequencing
4) identification of monoclones
5) screening of monoclones
6) lead optimization

A

quality control of starting library

B

improved analysis of enriched libraries

C

lead optimization

- accelerated and improved drug discovery and diagnostics
- advanced patenting strategies

* CoMmon PAtternS
Quality control of starting library

- Reliable quality information on the basis of millions of sequences.
  - Check of random region length,
  - Check of random or user defined distribution of building blocks (aa, codons)
  - Actual vs. target performance of building block usage
  - Check for contaminations
- High quality information allows the optimization of starting libraries.
Check of random region length at high resolution

A good starting library: random region length distribution of almost all 3.8 Mio. sequences reflects the library design.

Zooming allows to analyze down to a resolution of a couple of sequences. Traces of ligands with different loop sizes get apparent.
Check for distribution of building blocks

Overview of amino acid distribution enables to assess if synthesis measures up to your expectation.
Comparison of actual vs. target performance

Actual vs. target performance (loop 2)

- Share of individual amino acids [%]
- Actual vs. Target performance

loop 2 (9aa)
Identification of sequences with higher copy number

Check for Contaminations

- Histogram visualizes the frequency distribution of full length sequences
- Sequences can be checked against existing databases to identify already known clones
Check for contaminations

Contaminations can be sequences known

- to bind tags (e.g. His-tags, Strep-tags, …)
- to interact unspecifically with surfaces (separation matrices)
- be replication parasites

✓ User defined databases enable the easy detection of such potentially harmful sequences early in starting and enriched libraries.
Analysis of enriched libraries

Improved analysis means:

- **Early identification** of ligands
- Identification of **rare ligands** (not accessible by the conventional approach)
- Identification of ligands in **the context of sequence families**
- Identification of ligands with the **expected binding characteristics**
  - Intelligent “*in silico* picking” of promising representatives
  - Comparing differentiating selection experiments on a monoclonal level

©2013 by AptalT GmbH
Comparison of selection experiments

Assessing and comparing the efficiency of selection experiments at high resolution

- frequency of individual clones can be assessed.
- number of evolving clones can be assessed at different time points.
- thereby the diversity of enriched populations can be considered at high resolution.

⇒ COMPAS is a useful tool for the optimization of selection conditions and the optimal adjustment of screening platforms.

©2014 by AptaIT GmbH
Deconvolution of enriched libraries or representative “in silico picking” of ligands

**Goal:** Pre-selection of clones for binding/functional tests, that cover as much as possible different binding epitopes on the target protein.

**Assumption:** Related sequences more likely address the same binding epitopes on the target protein than sequences having no similarity.

**Action:** Pre-selection of maximal different sequences as representatives for functional testing.
Deconvolution of enriched libraries or representative "in silico picking" of ligands

Deconvolution procedure by intelligent clustering of enriched sequences by two steps:

(1) Clustering into sequence families/patterns
(2) Clustering into sequence family clans

Deconvolution step 1

Population (enriched) → Family "Hundreds" → Clan "Tens" → Individual member "hit"

Deconvolution step 2
Grouping into sequence families

deconvolution step 1

- Select one candidate of each family for initial binding analysis.
**Choice of ligands by comparison of experiments**

*In silico* tracking of monoclonal sequences over different experiments

Goal: identification of ligands that bind the target in selection variant 1 and 2

- Ligand evolution can be analyzed on monoclonal level.
- Clone “4” binds target protein presented in two different ways.

> Clones with expected binding characteristics can be identified on the in silico level.

<table>
<thead>
<tr>
<th>Clone</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>LVCFRIAKWCTTFVFCRHS</td>
</tr>
<tr>
<td>3</td>
<td>IWCSVQNMQSFWMECTTFVFCMIE</td>
</tr>
<tr>
<td>4</td>
<td>IACISWKTVTCTTFVFCSYS</td>
</tr>
<tr>
<td>5</td>
<td>FQCIVTMSSCTTFVFCRIK</td>
</tr>
<tr>
<td>6</td>
<td>GICKGVIRGCTTFVFCWIE</td>
</tr>
<tr>
<td>7</td>
<td>RVCTTFVFCWIE</td>
</tr>
<tr>
<td>8</td>
<td>QFCVWFGARVGDCTTFVFCRNE</td>
</tr>
<tr>
<td>9</td>
<td>RSCIKWTKATFTCTTFVFCYTE</td>
</tr>
<tr>
<td>10</td>
<td>MTCLRATMTCTTFVCAIV</td>
</tr>
<tr>
<td>11</td>
<td>FICVANIGCTTFVFCWAV</td>
</tr>
<tr>
<td>12</td>
<td>GSCKMKSRLCTTFVFCRMK</td>
</tr>
<tr>
<td>13</td>
<td>FGCFTFKWSVTATCTTFVFCRYE</td>
</tr>
</tbody>
</table>

©2014 by AptaIT GmbH
Comparative analysis - identification of ligands addressing defined binding sites

**wet lab:** *in vitro* evolution experiment (SELEX, eight cycles)

- MP1 (wt) → MP1 (wt)
- MP1 (mt) → MP1 (mt)

**Cycle** 1 - 6: 7 → 8

**wet lab:** filter binding assay

**in silico:** COMPAS analysis

- Discriminatory binding of clone MP 1-4 could already be predicted on the *in silico* level.
- Selectivity for the wild-type protein could be confirmed in binding assay.

©2013 by AptaIT GmbH
Opening the black box of *in vitro* selection

Evaluation of enrichment behavior on the basis of enrichment fold values

- “Mouse click” on dot prints peptide info:

<table>
<thead>
<tr>
<th>Clone No</th>
<th>Sequence</th>
<th>Enrichment factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>LVCFRIA...</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>IACISWK...</td>
<td>10</td>
</tr>
<tr>
<td>62</td>
<td>SKCATVTL...</td>
<td>14</td>
</tr>
</tbody>
</table>

- Color ramp: enrichment fold values are visualized by colors:
  - high enrichment
  - low enrichment

- Ligand evolution (increase or decrease) can be assessed under the aspect of enrichment fold.

→ Response to evolution pressure can be evaluated on the level of individual clones.
Grouping into family clans

Deconvolution step 2

Heatmap highlighting clusters with peptide families with highest similarity

Consensus sequences (each representing a defined peptide family)

©2014 by AptaIT GmbH
Grouping into family clans

Zoom into cluster of a family clan:

- Heatmap visualizes a cluster of sequence families (clan)
- Families C19, C24, C7, C189, C28, … have some degree of similarity (%)

⇒ Number of sequences to be tested can be adapted to capacity of the binding format.

©2014 by AptaIT GmbH
Summary of main results

Overview of relative frequencies of identified peptides

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Selection variant 1</th>
<th>Selection variant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Round 1</td>
<td>Round 2</td>
</tr>
<tr>
<td>Clone 2</td>
<td>LVCFRIAKWCTTFVFCRHS</td>
<td>0,1390%</td>
<td>1,4050%</td>
</tr>
<tr>
<td>Clone 3</td>
<td>IWCSVQNMQSFWMECTTFVFCMIE</td>
<td>0,0004%</td>
<td>1,3090%</td>
</tr>
<tr>
<td>Clone 4</td>
<td>IACISWKTIVTCTTFVFCSYS</td>
<td>0,0331%</td>
<td>0,6421%</td>
</tr>
<tr>
<td>Clone 5</td>
<td>FQICVTMSCTTTFVFCKI</td>
<td>0,1027%</td>
<td>0,5791%</td>
</tr>
<tr>
<td>Clone 6</td>
<td>GICGKVIGRICTTTFVFCEWIE</td>
<td>0,0005%</td>
<td>0,5554%</td>
</tr>
<tr>
<td>Clone 7</td>
<td>RVCTTFVFCWIE</td>
<td>0,0117%</td>
<td>0,0140%</td>
</tr>
<tr>
<td>Clone 8</td>
<td>QFCVWFGARVKTCTTFVFCTRNE</td>
<td>0,0022%</td>
<td>0,4087%</td>
</tr>
<tr>
<td>Clone 9</td>
<td>RSCIKWTKATFTCTTFVFCTE</td>
<td>0,0863%</td>
<td>0,3723%</td>
</tr>
<tr>
<td>Clone 10</td>
<td>MTCLRATMTCTTFVFCAIV</td>
<td>0,0034%</td>
<td>0,3174%</td>
</tr>
<tr>
<td>Clone 11</td>
<td>FICVANIIGCTTTFVFCCWAV</td>
<td>0,0311%</td>
<td>0,2706%</td>
</tr>
<tr>
<td>Clone 12</td>
<td>GSCKMKSRLCTTTFVFCKM</td>
<td>0,0018%</td>
<td>0,2562%</td>
</tr>
<tr>
<td>Clone 13</td>
<td>FCGTFFKWSVTATCTTFVFRCYE</td>
<td>0,0850%</td>
<td>0,2396%</td>
</tr>
<tr>
<td>Clone 14</td>
<td>WMCFVAMRYCTTTFVFECFIE</td>
<td>0,0010%</td>
<td>0,2338%</td>
</tr>
<tr>
<td>Clone 150</td>
<td>WQCTFWKVTCTTTFVFCDMK</td>
<td>0,0048%</td>
<td>0,0453%</td>
</tr>
</tbody>
</table>

- Clones are given in the order of frequency.
- User defined formatting/sorting/filtering is possible to highlight results of interest.
### Summary of main results

**Overview of amplification fold values**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequenz</th>
<th>Selection variant 1</th>
<th>Selection variant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Round 1 -&gt; Round 2</td>
<td>Round 2 -&gt; Round 3</td>
</tr>
<tr>
<td>Clone 2</td>
<td>LVCFRIAKWCTTFVFCRHS</td>
<td>10,11</td>
<td>4,47</td>
</tr>
<tr>
<td>Clone 3</td>
<td>IWCSVQNQMQSFWMECTTFVFCMIE</td>
<td>3253,79</td>
<td>0,74</td>
</tr>
<tr>
<td>Clone 4</td>
<td>IACISWKTIVKCTTFVFCSYS</td>
<td>19,42</td>
<td>9,98</td>
</tr>
<tr>
<td>Clone 5</td>
<td>FQCIVTMSSCTTFVFCRIK</td>
<td>5,64</td>
<td>1,76</td>
</tr>
<tr>
<td>Clone 6</td>
<td>GICKGKVIGRIGCTTFVFCWIE</td>
<td>1060,33</td>
<td>0,71</td>
</tr>
<tr>
<td>Clone 7</td>
<td>RVCDFVFCWIE</td>
<td>1,20</td>
<td>0,92</td>
</tr>
<tr>
<td>Clone 8</td>
<td>QFCVWFGARVVDCTTFVCRNE</td>
<td>189,74</td>
<td>6,57</td>
</tr>
<tr>
<td>Clone 9</td>
<td>RSCIKWTKATFTCTTFVFCTCLE</td>
<td>4,32</td>
<td>2,06</td>
</tr>
<tr>
<td>Clone 10</td>
<td>MTCLRATMTCTTFVFCAIV</td>
<td>93,41</td>
<td>6,32</td>
</tr>
<tr>
<td>Clone 11</td>
<td>FICVANIIGCTTFVFCWAV</td>
<td>8,70</td>
<td>8,29</td>
</tr>
<tr>
<td>Clone 12</td>
<td>GSCKMKSRCLCTTFVFCRMK</td>
<td>143,85</td>
<td>0,98</td>
</tr>
<tr>
<td>Clone 49</td>
<td>LICSVSVFMASVKSCTTFVFCAIV</td>
<td></td>
<td>172,08</td>
</tr>
<tr>
<td>Clone 62</td>
<td>SKCATVTLICTTFVFCRIE</td>
<td>329,35</td>
<td>13,95</td>
</tr>
<tr>
<td>Clone 150</td>
<td>WQCTFWKVTCTTFVFCDMK</td>
<td>9,40</td>
<td>1,01</td>
</tr>
</tbody>
</table>

- Clones are given in the order of frequency.
- User defined formatting/sorting filtering is possible to enable the identification of monoclonal clones that respond to the set evolution pressure.
Optimization

- Having a closer look on defined sequence families
- Identification of relevant sequence positions

©2013 by AptalT GmbH
Secondary screening of relevant sequence families

- Relevant clones identified
- Focus on relevant family and clan members
- Focused wet lab analysis of sequences covering this relevant sequence space

©2013 by AptalIT GmbH
Having a closer look on defined families:

Identification of constant and variable regions

- Conserved as well as variable sequence positions can be identified for sequence families
- COMPAS gives additional information to which extent nucleotides/amino acids can be replaced against others.
- Already information on the in silico level that would have to be generated otherwise in laborious doped selection experiments in conventional approaches.
Having a closer look on defined family clans:

Identification of minimal binding motifs

Common conserved and variable sequence positions can be identified for related sequence families
The integration of NGS and COMPAS analysis into the in vitro selection workflow enables...

✓ ...quality control of starting libraries.
  • Diversity
  • Distribution of NTs/AAs over sequence positions
  • Actual vs. target performance of building blocks

✓ ...improved analysis of enriched libraries.
  • Ligand prediction in the first enrichment cycles
  • Identification of rare ligands (not accessible by conventional cloning and sequencing)
  • Prediction of ligands addressing defined binding sites
  • Clustering of ligands in families and clans for a better target epitope coverage

✓ ...lead optimization and advanced patenting strategies.
  • Comprehensive information of conserved and variable sequence regions at high resolution
Thank you!

AptaIT team:
Dr. Carsten Gröber (Bioinformatics)
Christian Grohmann (Bioinformatics)
Dr. Raymund Buhmann (CEO)
Dr. Michael Blank (CSO)
Dr. Michael Blind (Representative)