Knowledge-driven De Novo Drug Design (Focusing Multi-Objective Optimization Search)

20090720
Leiden, The Netherlands

C. A. Nicolaou\textsuperscript{1, 2} C. S. Pattichis\textsuperscript{3} C. Kannas\textsuperscript{2}

\textsuperscript{1} CaSToRC, The Cyprus Institute, P.O.Box 27456, CY-1645 Nicosia, Cyprus
\textsuperscript{2} Noesis Chemoinformatics, Metochiou 66, CY-1599 Nicosia, Cyprus
\textsuperscript{3} Computer Science Department, University of Cyprus P.O.Box.20537, CY-1678 Nicosia, Cyprus
Outline

• Background
  – De Novo Design
  – Drug Discovery and Optimization
• Current status – Available information
  – Data, Knowledge
  – Methods
• Our contribution (to date…)
• Issues and Considerations
  – Providing for the unknown
  – Moving to the HPC world
• Concluding remarks
De Novo Drug (DND) design is currently experiencing a resurgence
   – Publications, applications, success stories, … [Schneider09]
Techniques with improved reported performance
Ongoing research on…
   – Addressing the synthesizability concern for the proposed compounds (partly)
   – Identifying suitable algorithmic methods for compound design and optimization
   – Proposing appropriate building blocks for the virtual synthesis of compounds
   – Measuring the performance of the designed compounds
Recognizing the multi-objective nature of drugs [Nicolaou07]
   – Somehow taken into account (sometimes)
   – Multiobjective-optimization algorithms
A Closer Look at Drug Discovery

• Drug Discovery is by nature a multi-objective problem
  – Potency, ADME, Toxicity, Synthesizability, Cost, ...
• Drugs are compromises among the various objectives
• Modern drug discovery process largely follows a linear paradigm
• However, models, information, data that can lead to the definition and exploitation of multiple pharmaceutically relevant objectives exist
  – Results from various screening experiments running in parallel
  – ADME-Toxicity physical & computational models
• Plenty of room to:
  – Exploit available information/knowledge
  – Apply Multiobjective Optimization (MOOP) technology
De Novo Drug Design

- Designing drug-like molecules
  - Satisfying a collection of, often conflicting, criteria
- Finding solutions to large multi-objective combinatorial problems
  - Estimated $10^{60}$ chemical structures using C, N, O, S [Bohacek96]
  - Recent work on enumerating all possible organic molecules of up to 13 atoms containing C, N, O, S, Cl, under some constraints (simple valency, chemical stability, synthetic feasibility) produced GDB-13 with 970 million compounds [Blum09]
- An exhaustive search is difficult/impossible
Existing DND Approaches

- Handling objectives
  - Overwhelming majority - single objective
  - Multiple objectives often averaged
    - Problem transformed a’priori to single objective
- Often impose constraints to limit the search space
- Design molecules:
  - Fitting into a well known receptor
  - Resembling a known drug/ligand
- Employ various search methods
  - Overwhelmingly use EA-based approach
- Often produce molecules with severe limitations in practice
  - lack of consideration of other pharmaceutically relevant properties
EA-based De Novo Design Methods

- Several applications in literature (~20) since the early 90’s
- Genotype Encoding:
  - Genes: 3-to-1 use fragments over atoms/bonds
  - Chromosome representation:
    - Evenly split between linear, tree-structure, graph
- EA type:
  - Genetic Algorithms (often with constrained crossover)
  - Evolutionary Strategies
- Problems/limitations with respect to rings
  - Response: no crossover, ring elimination via “special” nodes etc.
  - Graph-based chromosome methods partly tackle the problem (some)
- Primary constraint evenly split between ligand and receptor-based
- Substantial research in chemo- side of methodology
  - Info- side (algorithmic) of methodology still developing
Multi-Objective EA (MOEA)

• Extension of EAs; address multiple objectives simultaneously
• Additional components:
  – Fitness assessment of each individual solution to all objectives
  – Pareto ranking [establishment of domination relations]
  – Efficiency calculation and selection based on the rank.

• Pros:
  – Population-based approach
  – Simultaneous search of multiple search space regions
  – Identification of numerous Pareto solutions in a single run.
  – No constrains on the morphology of the search space
  – Suitable for complex, multi-modal surfaces [MOOP problems]
MOEA outline

Generate initial population $P$
Evaluate solutions in $P$ against objectives $O_{1-n}$
Assign Pareto-rank to solutions
Assign efficiency value to solutions based on Pareto-rank
While Not Stop Condition:
  Select parents $P_{parents}$ in proportion to efficiency values
  Generate population $P_{offspring}$ by reproduction of $P_{parents}$
    Mutation on individual parents
    Crossover on pairs of parents
  Evaluate solutions in $P_{offspring}$ against objectives $O_{1-n}$
  Merge $P$, $P_{offspring}$ to create $P_{new}$
  Assign Pareto-rank to solutions
  Assign efficiency value to solutions based on Pareto-rank
  Select $P_{new}$ from $P$, $P_{offspring}$

[Fonseca, 1993]
Open Issues

• Choosing a specific optimization algorithm
  – Possible adaptations to better suite molecular structure design
• Exploiting existing, problem specific knowledge to guide the optimization process
• Improving the scalability of the proposed methods to cope with the immense chemical space of possible, drug-like compounds
• ...
MOOP Research
(from the info-side)

- Niching
  - Preserve population diversity
- Elitism
  - Avoid compromise solution loss
  - Pareto-archive
- Exploit available information
  - Reuse discovered knowledge
Available Information

• Data...
  – Public
    • Protein Data Bank
      – Repository for the 3-D structural data of large biological molecules
      – >58k entries (June 2009)
    • PubChem Data Source Information
      – Contains small molecules (<1000 atoms & <1000 bonds)
      – Over 80 sources of information
      – 1644 BioAssay, (bioactivity results) from high-throughput screening programs with several million values
      – 37 million compound entries
  – Private, corporate, proprietary, ...

• Knowledge...
  – Drugs, patents, publications, models, ...
Memetic Algorithms (MA)

• Hybrid Algorithms:
  – EAs + a stage of individual optimization or learning
    • Genetic-Local Search
  – Incorporating domain knowledge into EA driven global search
• “Memes” vs Genes
  – A “meme” stands for “unit of imitation” in cultural transmission
  – “The selfish gene” [Dawkins76]
• Pros: Faster, substantially better performance in several applications
• Cons [open research issues]:
  – when to use local search;
  – which individuals to improve using local search
  – at what intensity to apply local search
  – how to integrate genetic operators with local search
  – ...
MA outline

Generate initial population $P$
Evaluate solutions in $P$ against objective $O$
While Not Stop Condition:
  - Select parents $P_{parents}$ in proportion to fitness scores
  - Generate population $P_{offspring}$ by reproduction of $P_{parents}$
    - Mutation on individual parents
    - Crossover on pairs of parents
  - Select $P_{improve}$ subset:
    - Improve via local search
  - Evaluate solutions in $P_{offspring}$ against objective $O$
  - Select population $P_{new}$ from $P$, $P_{offspring}$

- Improvement via local search may take place using any search method including hill-climbing methods, Monte-Carlo, or even an EA.

[Krasnogor02]
Contribution

• Multi-Objective Optimization Method [Nicolaou09]
  – Knowledge-driven
  – Designed for optimal graph design/de novo design

• Key elements
  – methods for obtaining molecular building blocks
  – evolving the chemical graphs
  – scoring the designed molecules
  – exploiting domain specific knowledge.
Multi-objective Evolutionary Graphs Algorithm

• MEGA
  – Multi-objective
    • Evolutionary Algorithm
    • Pareto approach
  – Graph representation
    • Chromosome = Graph
    • Avoid information loss
  – Incorporates domain specific knowledge
    • Learn adaptively
    • Avoid random walks through space
    • Information rich fragments

1. Population of candidate molecules
2. Evolve molecular graphs
3. Compute fitness of each candidate to all objectives
4. Identify compromise candidates
5. Select representative candidates based on molecular structure
6. START
Obtaining Building Blocks

• Molecular Fragments
  – Several methods, i.e. Frameworks Analysis, Substructure Mining, RECAP, ...
  – Information rich
    • Attachment points
    • Attached group type...
  – Profiled to assess “privileged” status/weight
  – Adjustable property

• Atoms, Bonds

Substructure Mining – MCS-based

[Nicolaou & Pattichis, 2006]
Objectives - Measuring Fitness

• Molecular Scorers
  – Drug-likeness scorers
    • Complexity, Flexibility, MW, ...
  – Ligand-based scorers
    • Similarity measure
    • Requires target molecule(s)
  – Target-based scorers
    • Predicted binding affinity – Docking
    • Requires target structure

• Modular; Plug and run...
  – List of scores for each molecule
Ligand-based Objectives

• Given collection of known ligands...
• Design “similar” (or dissimilar) molecules
• Methods:
  – Descriptor-based
    • E.g 2D atom-pair, tanimoto similarity
  – Graph-based
    • Maximum Common Subgraph-based
  – Shape-based
    • Alignment of molecular metagraphs
  – QSAR model-based
Reduced graph mapping (Fuzzee)

- Molecular graph
- ... is reduced to metagraph
- ... compared to other molecules
- ... via an optimal weighted mapping solution
- Takes approximate topology and physicochemical properties into account
- Calculates tanimoto-like similarity on matching
- Approximately 1000 comparisons/s

www.chil2.de
Target-based Objectives

• Given collection of known, well-defined target receptors...
• Design molecules with high (or low) predicted binding affinity
• Methods:
  – Chil\(^2\) Glamdock
Docking with GlamDock

- Prepare the binding site by placing ideally interacting probes (NH, CO, 6-rings)
- Use these probes to place the ligand into binding site
- Monte carlo with basin hopping
  - Differentiable empirical scoring function
  - ChemScore like
  - Fitted to binding affinities
- Glamdock has been validated on several benchmarks
  - Kellenberger data set
  - Sc-PDB (>6000 complexes)

[Tietze & Apostolakis 07]
Hard Filtering

• Ranges of acceptable values for objective scores
• Rejection of individuals with scores out of one or more ranges
• E.g.
  – Molecular_weight = [300,800]
Pareto-rank

• Identify domination relations in each pair of compounds in the population
• For each solution s:
  – Count number n of solutions dominating it
  – Set solution rank to n+1
Efficiency Calculation

• Method A:
  – Proportional to solution rank
  – Issue: Niching
    • Domination of the population by one solution/chemotype
    • Better at some point of the evolution

• Method B:
  – Diversity Analysis
    • Graph (molecule) clustering
      – Current implementation uses Ward´s agglomerative method
    • Cluster profiling
  – Individual efficiency calculation based on Pareto-rank and cluster placement
Pareto-archive

• Incorporate elitism mechanism
• Elite population
  – An archive of all Pareto solutions in each iteration
  – External population, unlimited size***
• Requires maintenance
  – After new solution addition, re-rank and eliminate solutions found to be dominated
• Organize in clusters

[Zitzler et. al., 1999]
Parent Selection

- Basic EA methods
  - Best, Roulette, Tournament

- Population:
  - Current iteration population
  - Combination of current and elite population (if available)

- Using solution efficiency value
  - Clustering assignment (if available)
  - Success record of molecules from the cluster

- Subpopulations...
  - Formed by clusters of non-dominated solutions
  - Allow local exploration
Sample Results - 1

• Goal
  – Designing Selective Inhibitors
    • ER-β (2fzs1) Vs ER-α (1xpc)
    • Similarity to Tamoxifen

• MolScorers:
  – Glamdock score to ER-β, ER-α (inverse)
  – Descriptor-based distance to Tamoxifen
    • Used as a filter

• Initial population
  – Random compounds

• Fragment pool
  – Calculated from a small set of known ER inhibitors

• Multiple runs, variety of MEGA settings
  – Population size, number of generations, evolution settings
Indicative Results

• Compound 2098_1057
• Overlapped with Tamoxifen as co-crystallized in ER-β
Compound 2098_1057

ER-α docking position
More results
The Case of 3414_m3p6

ER-β docking position

ER-α docking “position”
Compromise Surface – 2D

- Actual compromise surface formed over several generations
- Gradual conversion
- Initial rounds produced nearly identical docking scores
- Later rounds produced several individuals with substantial difference in scores

y-axis score is inverted
Diversity – Elitism Effect

Pareto-front approximations obtained with eMEGA (left) at iterations 1, 10, 50, 100, 1000. Later generations have a more advanced (closer to the ideal point) and dense front approximation. A comparison between the Pareto-front obtained with MEGA and eMEGA. eMEGA front is more dense with larger spread.
Sample Results - 2

- **Goal**
  - Designing Selective Inhibitors (ER-β Vs ER-α)

- **MolScorers:**
  - Shape similarity to:
    - ER-α (2 known selective ligands),
    - ER-β (5 known selective ligands) (inverse)
  - Rule of 5: Used as a hard filter

- **Initial population**
  - Random compounds

- **Fragment pool**
  - Using Pubchem bioassays 629, 633 on ER-β, ER-α
  - Weighted fragments

- **Multiple runs, variety of MEGA settings**
  - Population size, number of generations, evolution settings
Performance Assessment

- **Issues**
  - Approximation to true Pareto front
  - Spread

- **Hypervolume**
  - Measures volume defined by Pareto-front approximation and a given, reference point
eMEGA

De Novo Molecular Design

Pareto Front

Performance Assessment - Hypervolume

Iterations
Algorithm Comparison - 1

Performance Assessment

- Hypervolume
- Epsilon Unary

Algorithms

moea_eMEGA  moea_MEGA  moea_MOGA
Algorithm Comparison - 2

Performance Assessment

Hypervolume

moea_eMEGA  moea_LSPEA  moea_MEGA  moea_MOGA

Epsilon Unary

moea_eMEGA  moea_LSPEA  moea_MEGA  moea_MOGA
Issues

• Using available knowledge appropriately
  – Allowing for the unknown to take place

• Moving into High-Performance Computing
  – Exploiting modern raw processing power
The Black Swan Effect

- Knowledge is good
  - But how much do we really know....
    - Turkey longevity model...
    - QSAR...
- Black swans
  - Surprise
  - Major impact
  - Rationalized a’posteriori
    (Taleb, 2007)

A Turkey is fed for a 1000 days. And then, on the 1001\textsuperscript{st} day, it has a surprise.

www.edge.org/3rd_culture/taleb08/taleb08_index.html
Coping with Black Swans

• There are limits to our knowledge...
• In the case of predictive modeling
  – Recognize regions outside model domain
  – Do not use the model in those regions...
• In the case of optimization??
  – “Avoid optimization. Love redundancy”
    • Naleb’s advice
  – Biological systems example
  – Provide for unexpected events!
• Ongoing work on evolutionary characteristics needed for Black Swan appearance
High Performance Computing (HPC) 

**Potential Impact**

- **Example:**
  - eMEGA run:
    - 200 population, 1000 iterations, high mutation, crossover probabilities
    - 120,000 molecules ($1.2 \times 10^5$)
  - Estimated $10^{60}$ (or $9.7 \times 10^8$) molecules
- **Gigaflop/s Vs Teraflop/s**
  - (Petaflop/s already here...)
- Exploring larger portion of the space
- Reducing execution time
- Handling more objectives/more complex space
Hardware Advances

- Processor technology advancement has changed drastically since 2003
- Core performance increase has slowed down
- Improvements observed due to “multi-core” architecture of modern processors
- Many-core coming...
- Current software development practices are based on a linear programming concept
- In order to exploit the new hardware architectures an application development paradigm shift is needed

Hoebling 2009
Supercomputer Architecture

- Modern supercomputers are massively parallel systems combining thousands of processors/cores and interconnect hardware to achieve a high cumulative performance.
- Main differences in:
  - Processor type/number
  - Number of cores per processor
  - Interconnect type
  - Degree of integration
- Use "off the shelf" components e.g. commercially available server-class microprocessors
  - costs of chip development and production make it uneconomical to design custom chips
JUGENE – Example System

- Site: Juelich Supercomputing Center
  - #3 in the Top500 list
  - Installation Year 2009
- System: IBM BlueGene (BlueGene/P)
- Operating System: SuSE Linux Enterprise
- Interconnect: Proprietary
- Processor PowerPC 450 850 MHz (3.4 GFlops)
- Cores: 294912;
- Performance(Gflops): 825500
HPC in Life Sciences

- Collection and maintenance of large collections of unstable and heterogeneous public datasets
- Data management for extremely diverse data types generated by multiple technologies, but requiring common references (ontologies, gene identifiers, normalization strategies etc)
- Ever-increasing need for simulations to assess the statistical significance and properties of observations and to validate models
- Most code was written by scientists without formal training in software engineering, and requires both verification and optimization
- A life science HPC center has to accommodate a large number of applications with very different behaviors
- Many life science applications can benefit from hardware acceleration, and this should be leveraged whenever possible

Jongeneel 2008
HPC in Life Sciences

• Rather slow adoption...
• Issues
  – Parallel architectures of supercomputers dictate the use of special programming techniques to exploit their speed.
    • Need of special libraries to share data between nodes.
    • Lack of experts
  – Significant effort is required to optimize a problem for the interconnect characteristics of the machine
  – Different applications requiring different optimization techniques
    • I/O bound vs CPU bound
    • Time-critical vs batch
    • Short (< 1 min) and long (> 1 month) execution times
    • Conflicting user requirements and fair share issues
  – Licensing... (Big topic. Period.)
Ongoing Work

• Cy-Tera: A new research infrastructure to host a high-performance computing system:
  – Minimum of 20 Tflop/s peak performance and 100 Tbytes of storage
  – Deployment: mid 2010

• Major Goals
  – Support research activities in computational science, scientific computing

• Projects
  – Porting to and optimization of existing code for new multi-core processors
  – Adaptation of applications to cluster/parallel environments
  – Ad hoc software development for technology platforms
  – Use of special accelerators for compute-intensive algorithms
Concluding Remarks

- Drug discovery is inherently multi-objective
- Nowadays, there exist MOOP methodologies that can be applied to speed up drug design and optimization
- Knowledge exploitation
  – Early definition of additional objectives is feasible; models and data are there
- MEGA or similar algorithms can assist
  – Combine global search abilities of the EA’s with detailed local search via the incorporation of domain specific knowledge
- Our results show that domain specific knowledge can have a positive impact
  • should in principle be exploited
- Recognize the limits of our knowledge
  – And provide for the unknown...
- Harnessing today’s HPC system power is feasible;
  – substantial investment needed in algorithm customization, porting applications to modern hardware architectures
  – The rewards will be substantial...
Acknowledgments

• Chil\textsubscript{2} Team
  – Joannis Apostolakis
  – LMU, Munich

• University of Athens
  – Med. Chemistry Dept
  – Prof. Manos Mikros

• Cyprus Institute
  – Prof. C. V. Jongeneel

• Noesis
  – Christos Kannas

• University of Cyprus
  – Medical Informatics Research Laboratory
Thanks for your attention!

Christos Nicolaou
- Noesis Chemoinformatics
- Computation-based Science and Technology Center
  The Cyprus Institute
  cnicolaou@cs.ucy.ac.cy