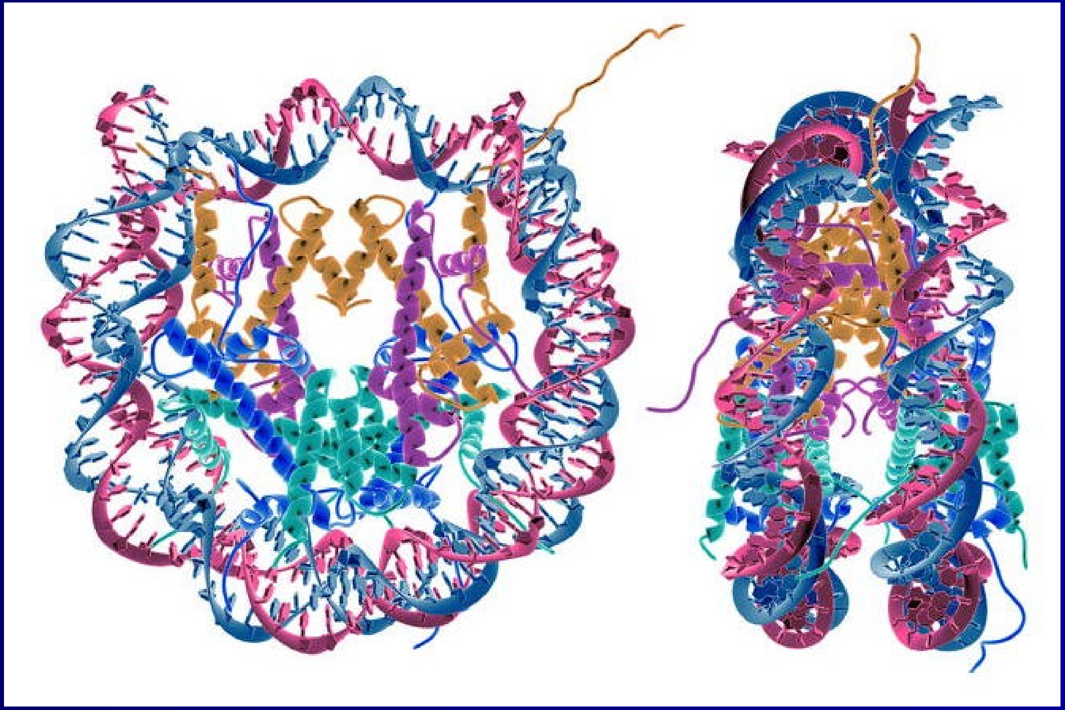
Welcome



Topics to be discussed

Background

Related Work

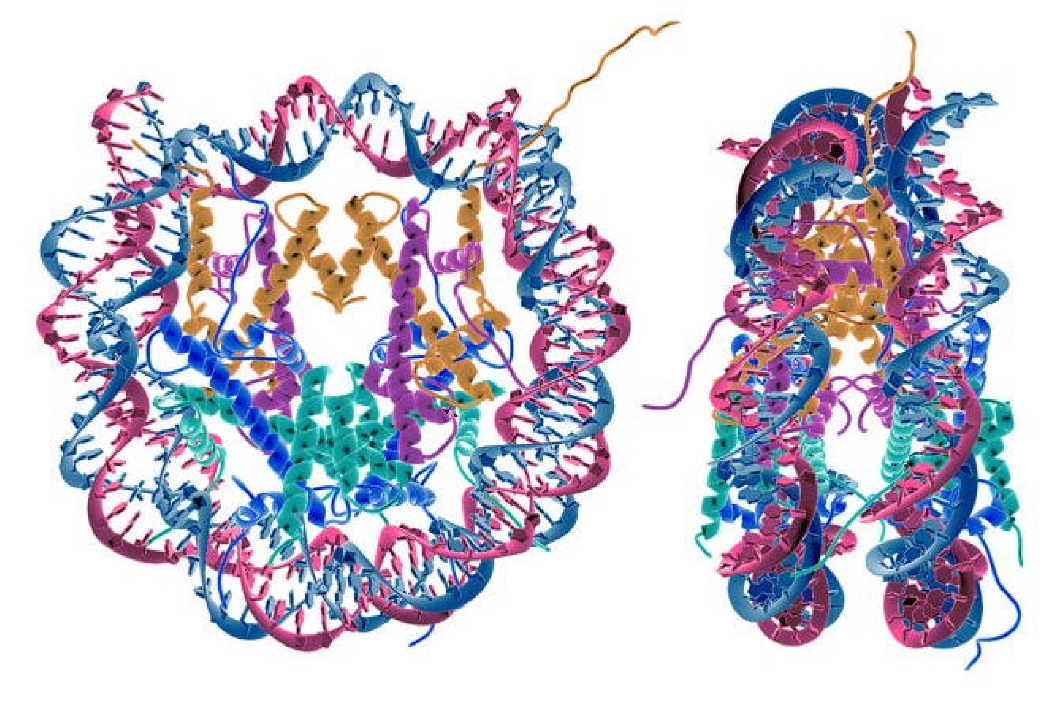
Problem and Proposed Solution

The Framework

Results and Applications

Conclusion

Nucleosome Structure



- Eight histone chains form the core protein octamer: a tetramer of (H3)2(H4)2, and two flanking H2A-H2B dimers.
- Histone chain central folds contact DNA superhelix.
- Amino-terminal tails extend beyond the core to form larger chromatin structures for packing.

- ~146 DNA base pairs wrap the octamer 1.65 times in a left-handed direction.
- Nucleosomes occur generally every 157-240 DNA base pairs.

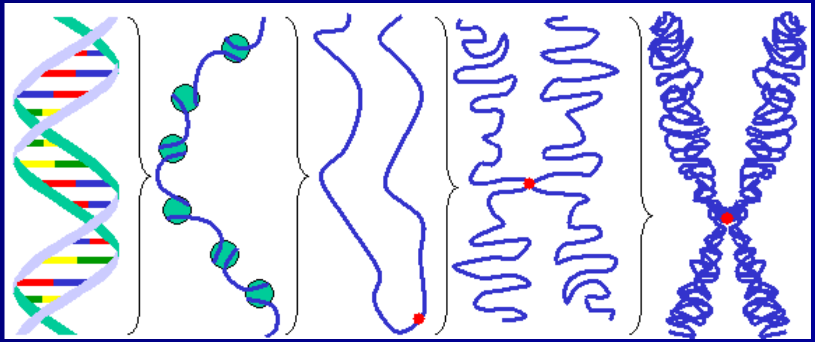
Evolutionarily Conserved

- The histone chains comprising the nucleosome core protein are some of the most conserved in nature (Kimball 2006).
- Arginine side chains in 12 of 14 DNA-histone binding positions are absolutely conserved among the major type histones of all species (Muthurajan et al., 2003).
- The histone chain H4 in the calf differs from H4 in the pea plant by only two amino acid residues out of the chain’s 102 amino acids.
- Allows understanding biological significance of nucleosomes by looking at a specific one, such as Xenopus laevis.
- Highlights emphasis on DNA sequence by mitigating the complexity introduced by differences in histones.

Chromatin Packaging

- Replication of the genetic material undergoes considerable packing of chromatin during prophase of mitosis; proper chromosomal division requires it.

- For the human diploid chromosome count of 46, the unpackaged DNA length would be over two meters. The ultimate size manageable for replication is just ten micrometers (Kimball 2006).

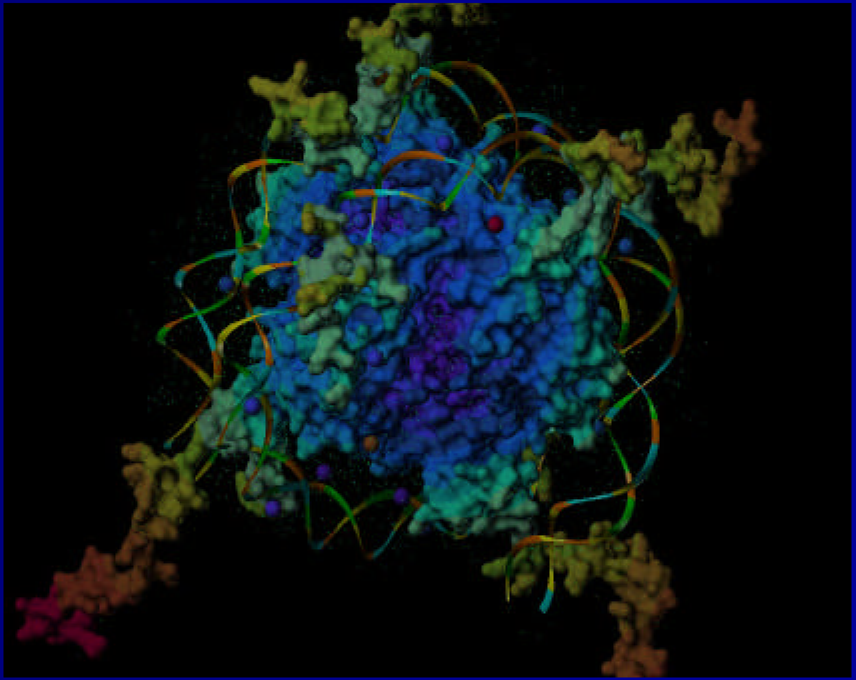


Transcription Regulation

- Proteins can locate unbound promoting regions of DNA and systematically dissociate neighboring nucleosomal DNA for processing.
- Positions of DNA sequence in the bound state could be recognized by proteins evolved to locate such sequences in the bound state.
- Nucleosomes can repress transcription by preventing the initiation of RNA polymerases, and thus preventing access to genes hidden in the bound configuration.

DNA Positioning

- Electrostatic interactions and hydrogen bonds bind DNA to the protein core.
- The surrounding environment of solvent, ions, and other proteins affect the binding state.
- There is strong interplay between DNA sequence and nucleosome stability.
- Relative binding affinity equals relative equilibrium stability.



Central to the principle of nucleosome positioning is energy minima.

DNA Sequence Effects

- Why does positioning of nucleosomes on ~146 DNA base pairs seem to occur preferentially? What are the sequence characteristics at regions conferring higher binding affinity?

- Nucleotides do not directly relate to binding affinity because only the backbone contacts with the protein core.

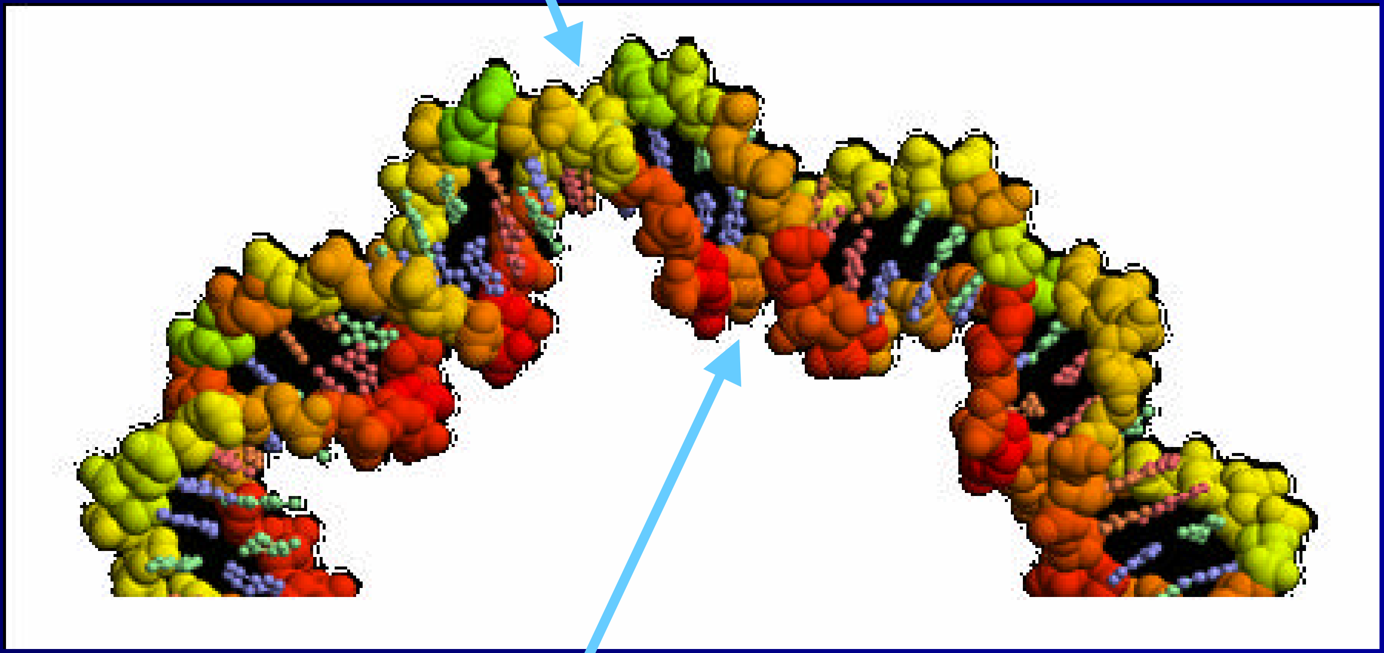
- All DNA has a degree of attractive nature to the core protein, some having a relatively higher binding affinity (1000-fold; Widom, 2001).

- Paradoxically, the nucleosome must form with most all sequences of DNA for chromatin folding and function, but in a way to permit eventual unfolding to expose the DNA.

DNA Sequence Effects, cont.

- Sequence specificity relies on one seemingly dominant feature of DNA: anisotropic deformability (bendability).
- Sequences possessing a high degree of bendability confer a higher likelihood of binding stable nucleosomes; sequence areas of poor bendability and curvature reflect positions with reduced likelihood.
- Which sequence effects support bendability and curvature?

DNA Bendability



Expansion site of the minor groove

Compression site of the minor groove

DNA Bendability, cont.

- Certain nucleotide arrangements must permit DNA curvature
- A manageable bending force is needed to form a wrapped state.
- What nucleotides in a sequence are needed to confer appropriate curvature?
- Where are interesting positions of bendability for nucleosome binding?

Topics to be discussed

- Background
- Related Work
- Problem and Proposed Solution
- The Framework
- Results and Applications
- Conclusion

Statistical Sequencing

- Statistical sequencing (Satchwell et al., 1986) was used to investigate sequence similarities in nucleosome forming DNA.
- Pattern: nucleotides in phase with the period of approximately 10.2 DNA base pairs.
  - short runs, two and three, of (A, T) nucleotides are positioned facing the octamer on the minor groove of DNA
  - similar runs of (G, C) face outward on the minor groove, away from the core.
- Fourier analysis of 177 statistically sequenced strands confirmed periodic waveforms of GpC and ApA occurrences around the same period.

Reconstitution Experiments

- Salt gradient dialysis is commonly used to reconstitute core proteins with DNA sequences in competition with one another.
- A subsequent quantitative analysis shows relative sequence preferences.
- Most all analysis shows directional bending preferences for (A,T)- and (G,C)-rich sequences.
- This method has populated many databases of well-forming nucleosomal DNA sequences.

Sequence Alignment

- A multiple sequence alignment procedure lines up common features using heuristics (Ioshikhes, 1996; Bolshoy et al., 1997).
- The procedure found patterns related to the frequency and position of base pairs.
  - The AA pattern was seen on the octamer facing minor groove.
  - The complementary TT was discovered six DNA base pairs down sequence in a symmetric assembly.
- Runs of AAA, found by Widlund et al. (1997), were six times more likely than the statistical expectation.

Topics to be discussed

Background

Related Work

<u>Problem and Proposed Solution</u>

The Framework

Results and Applications

Conclusion

Limitations of Current Methods

- Selection and reconstitution experiments explore an infinitesimal fraction of the full possible sequence space.

- Numerical bendability models exist but are parameterized independent of position on nucleosome and for DNA free in solution.

- Experimental findings speak to the qualitative features of well-forming nucleosome sequences, but no quantitative measure yet exists to correctly predict highly-affine sequences.

- Competition based experiments cannot differentiate between sequences within a set of highly-affine sequences.

Complications to Current Methods

- The requirements for DNA curvature on the nucleosome appear to be nonuniform around the histone octamer.

  - Super-coiled DNA surrounds the pseudo dyad axis with about 10.5 bp per turn, and about 10.0 bp per turn to the sides of the central region (Gale & Smerdon, 1988).

  - The central 15 DNA base pairs may be completely unbent (Ioshikhes, 1996).

- Highly flexible DNA has been identified near chromosome breakpoint regions that are associated with chromatin disruptions (Goode et al., 1996).

Molecular Modeling

- Computational simulation and modeling would allow for a complete picture of nucleosome interactions.

- Directly calculate interesting energy terms such as the internal energy of the DNA, and elucidate free-energy surface of DNA-histone bound state.

- Large-scale sequence comparison

- Experimental design freedom to evaluate conformational suites DNA sequence effects such as motifs and positions.

- Can take advantage of rapidly emerging computer power and intensity.

Molecular Modeling, cont.

- Molecular dynamics and energy minimization engines
    - CHARMM
    - NAMD

- Molecular topologies and force fields
    - CHARMM27
    - OPLS

- Biopolymer visualization and mutation software
    - Sybyl
    - VMD

- Tools used to enhance the biological coherence of molecular systems removed from their native environments.
    - GROMACS

Molecular Modeling, cont.

- Molecular dynamics and energy minimization engines simulate molecular systems

- Provided to the engine are molecular coordinates

- Definitions of how each atom behaves according to their respective chemical properties (force fields).

- Instructions for evolving the system over time

  - Equations of motion are calculated based on repulsive and attractive forces that each atom has with its neighbors and the overall electrostatic environment.

- By simulating the molecular system long enough, statistical principles can be used to infer the overall mechanics of the underlying system.

Topics to be discussed

Background

Related Work

Problem and Proposed Solution

<u>The Framework</u>

Results and Applications

Conclusion

Framework Overview

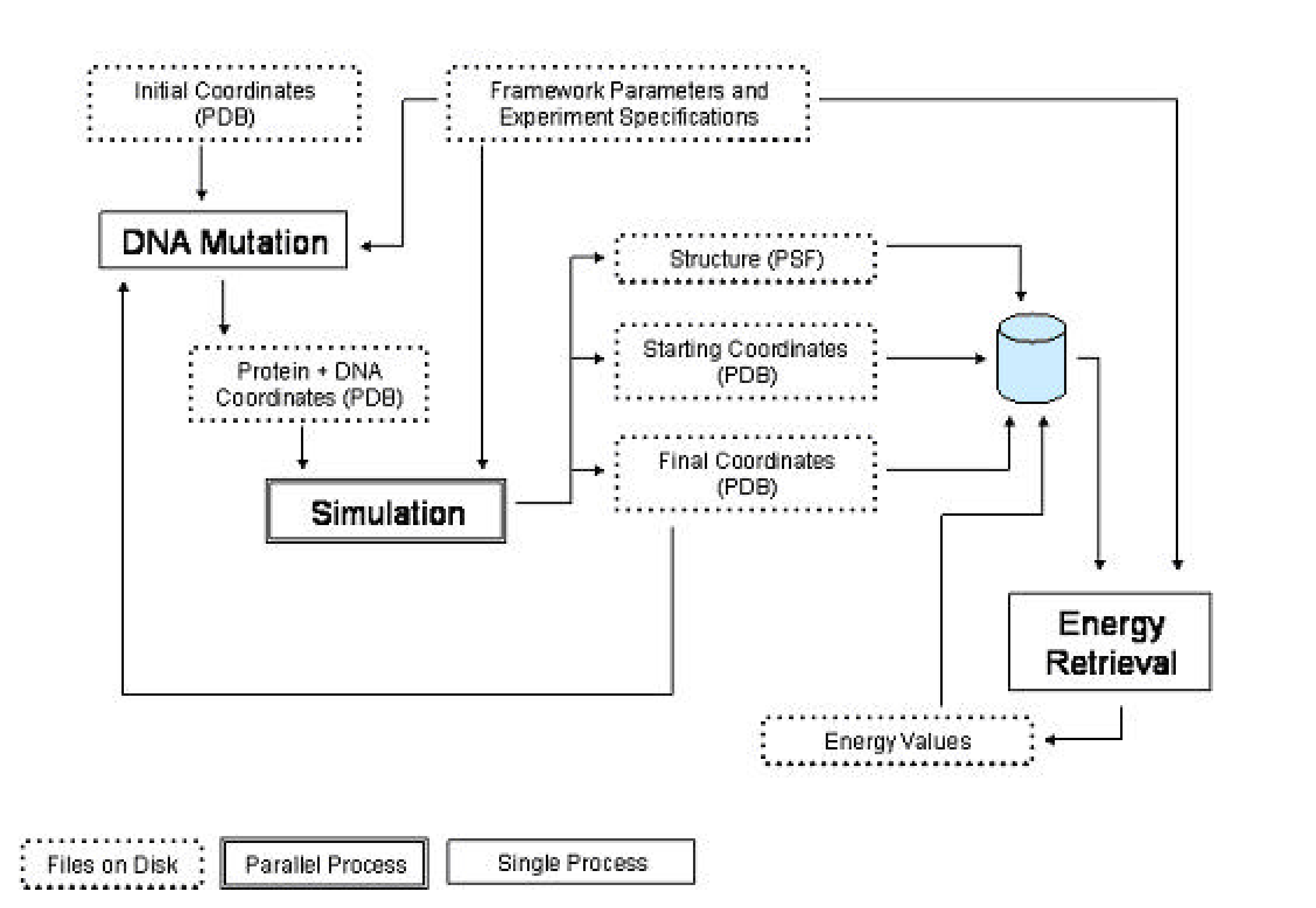
- The framework supports the computation of DNA sequence affinities using molecular modeling modules.

- The system provides for the definition of a suite of DNA molecules, which are automatically generated and simulated.

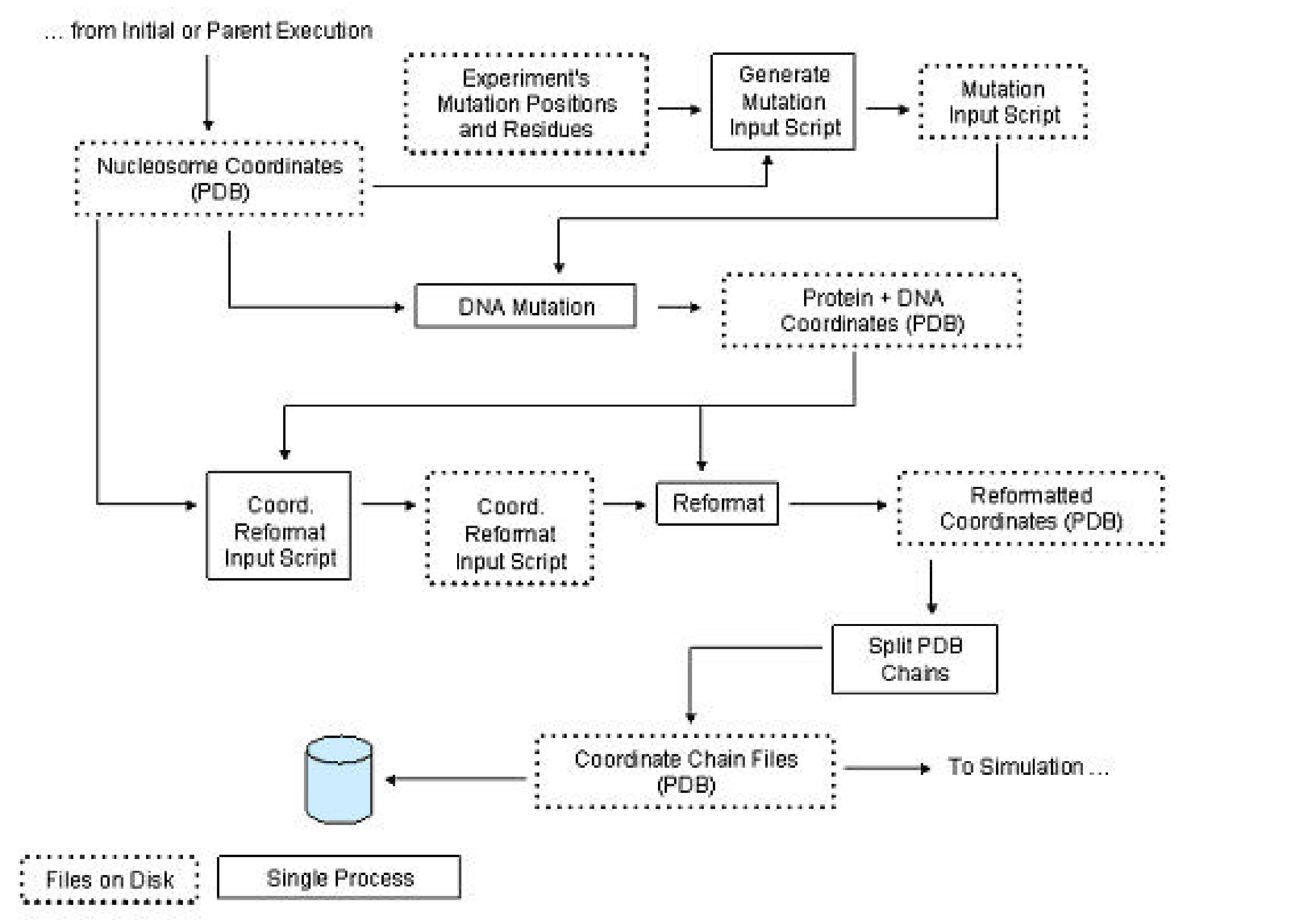
- The framework presented addresses computation requirements with parallel computing and by reducing processing requirements on the large number of DNA sequences through incremental nucleotide substitutions.

- The framework implements a heuristic for ordering the substitutions with stepwise energy relaxation on substituted DNA to minimize perturbations to the base structure.

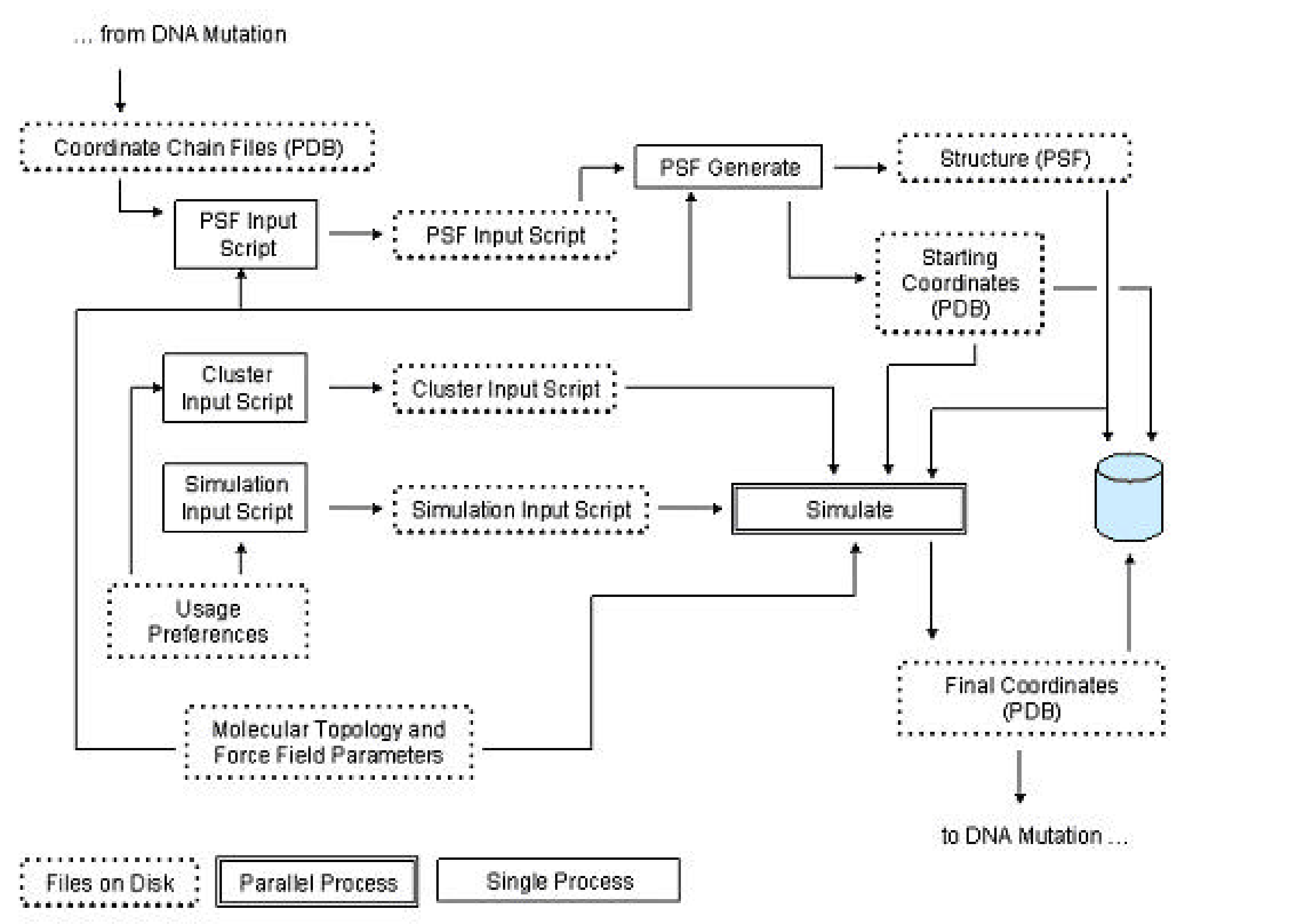
Framework Overview, cont.


Initial Coordinates (PDB)
Framework Parameters and Experiment Specifications
DNA Mutation
Structure (PSF)
Starting Coordinates (PDB)
Protein + DNA Coordinates (PDB)
Final Coordinates (PDB)
Simulation
Energy Retrieval
Energy Values
Files on Disk
Parallel Process
Single Process

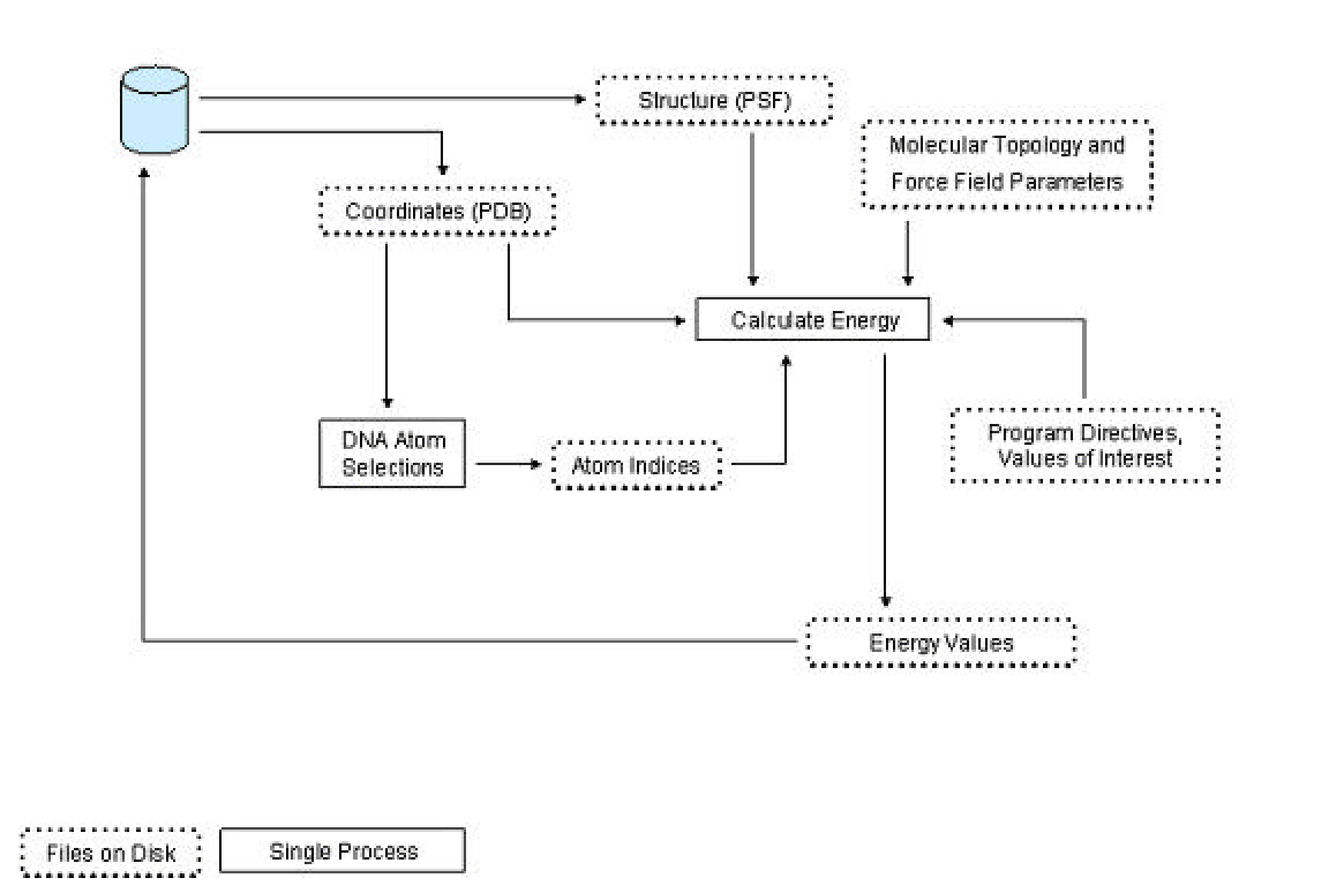
DNA Mutation


... from Initial or Parent Execution
Experiment's Mutation Positions and Residues
Generate Mutation Input Script
Mutation Input Script
Nucleosome Coordinates (PDB)
DNA Mutation
Protein + DNA Coordinates (PDB)
Coord. Reformat Input Script
Coord. Reformat Input Script
Reformat
Reformatted Coordinates (PDB)
Split PDB Chains
Coordinate Chain Files (PDB)
To Simulation ...
Files on Disk
Single Process

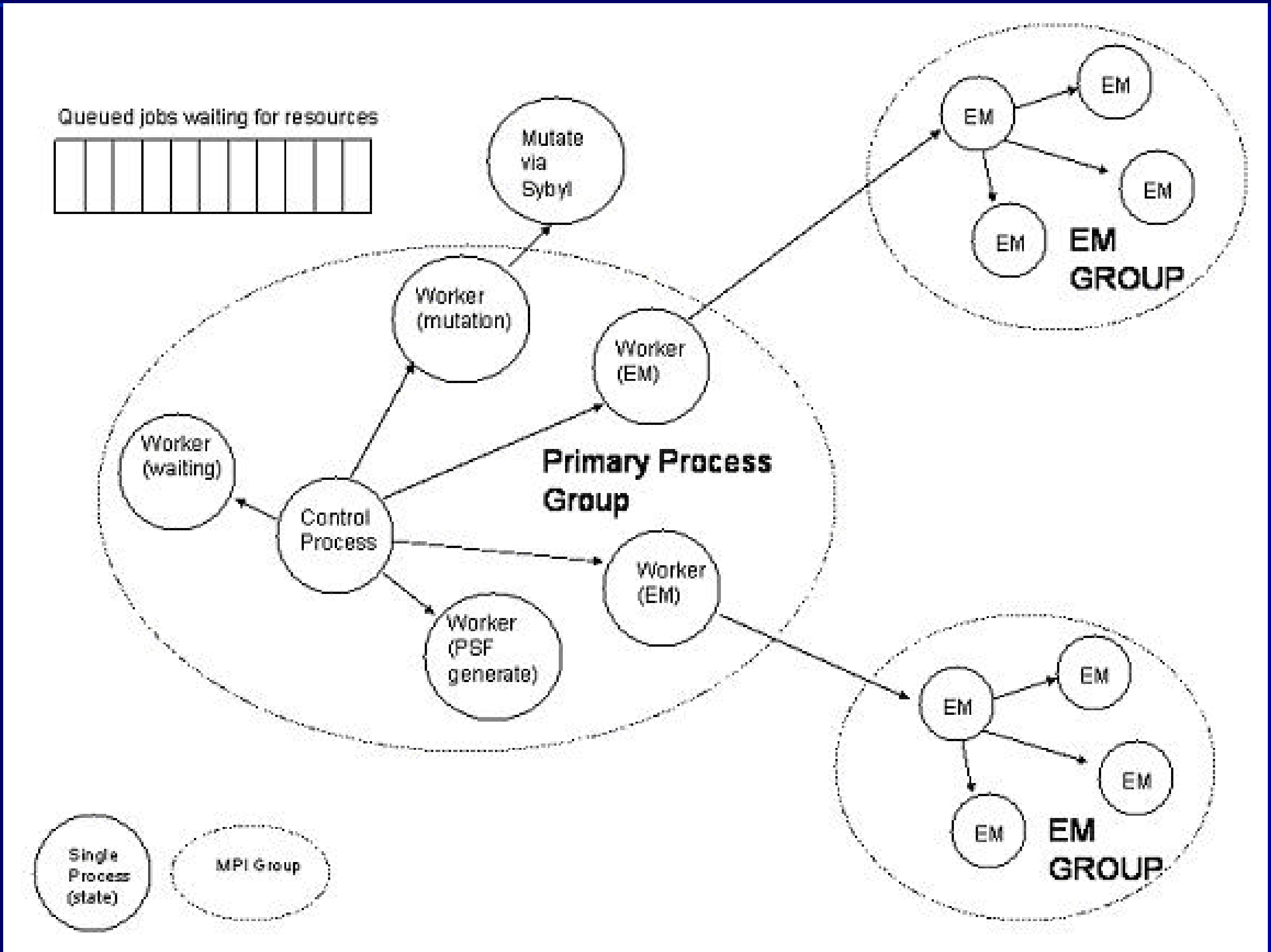
Simulation


... from DNA Mutation
Coordinate Chain Files (PDB)
PSF Input Script
PSF Input Script
PSF Generate
Structure (PSF)
Starting Coordinates (PDB)
Cluster Input Script
Cluster Input Script
Simulation Input Script
Simulation Input Script
Simulate
Usage Preferences
Molecular Topology and Force Field Parameters
Final Coordinates (PDB)
to DNA Mutation ...
Files on Disk
Parallel Process
Single Process

Energy Retrieval



Process View



Prototype Study

There are 12 sequence positions on either side of the central position of DNA sequence surrounding a nucleosome core considered important to nucleosome stability.

AAA tri-nucleotides contribute a significant degree to nucleosome stability when found +5, +15, +25, +35, +45, and +55 from the central DNA base pair position. Correspondingly, the tri-nucleotide contributes significantly at -5, -15, -25, -35, -45, and -55 from the center.

GGC tri-nucleotides contribute a significant degree to nucleosome stability when found +10, +20, +30, +40, +50, and +60 from the central DNA base pair position. Correspondingly, the tri-nucleotide contributes significantly at -10, -20, -30, -40, -50, and -60 from the center.

Prototype Study, cont.

Mutations are done symmetrically with respect to the DNA sequence center. In other words, a single configuration contains one tri-nucleotide motif at the same offset, both positive and negative from the center position.

Certain positions and combinations of a tri-nucleotide mutation may have a more pronounced affect on stability than others. For example, a single mutation of GGC at a central offset of 40 may contribute more to a lower energy configuration than the same mutation at central offset of 10.

Comparison of DNA sequence effects on nucleosome stability can be interpreted from differences in the total energy calculations of each engineered DNA conformation.

Framework Specifications

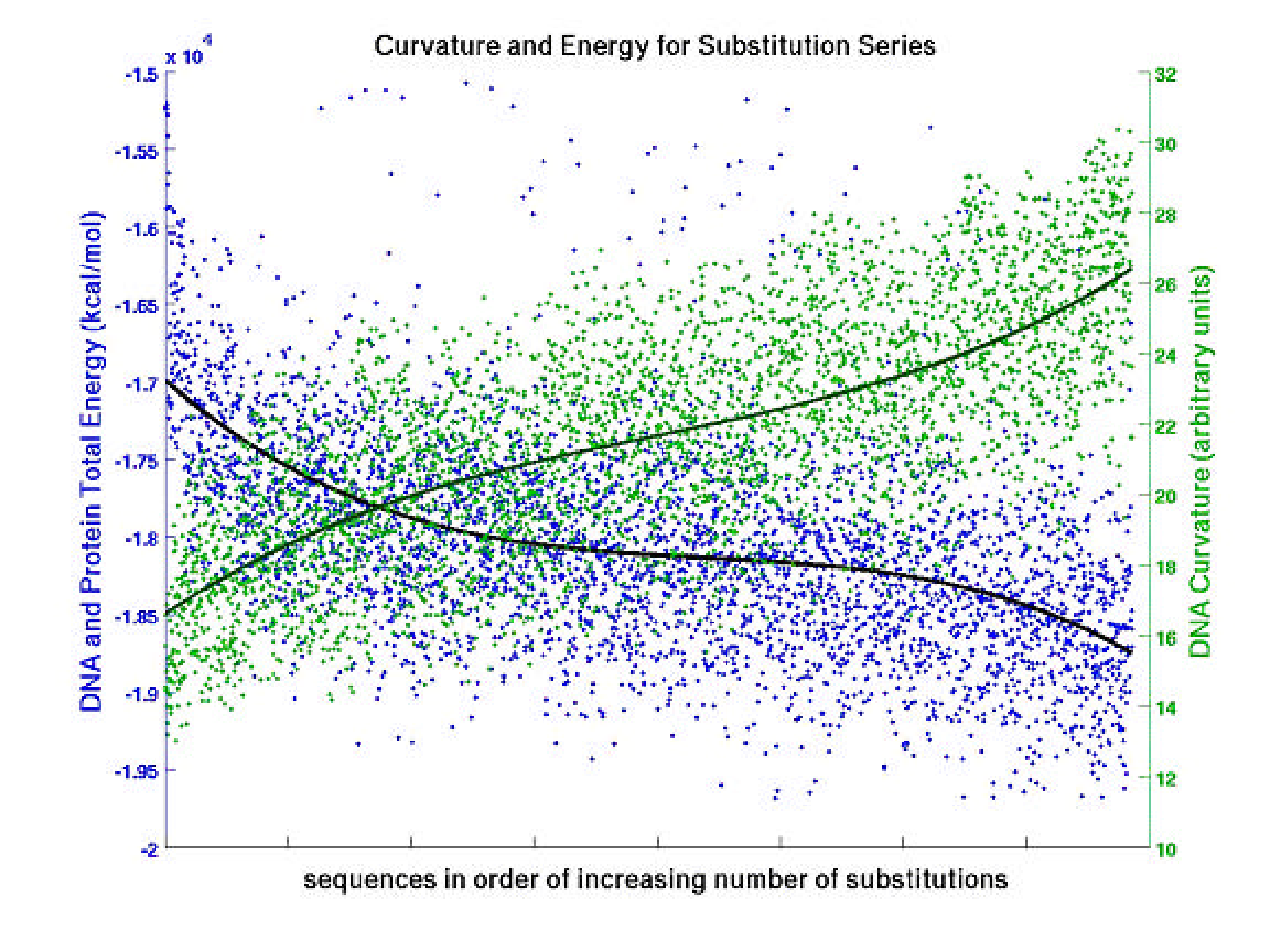
A series of thermalization and equilibration procedures of the nucleosome structure, 1KX5, were done in a fully solvated environment.

Mutation was done using the Biopolymer package of Sybyl using Secure Sockets Handling on the cluster.

Simulation was done using the CHARMM package with the CHARMM27 All-Hydrogen Protein and Nucleic Acid topology and parameters. Energy Minimization was done for 500 steps using implicit solvent on four processors.

Total energy was extracted from the suite of 4096 substitutions and plotted against the bendability assessment program, BEND.

Results


Curvature and Energy for Substitution Series
x 10^4
-1.5
-1.55
-1.6
-1.65
-1.7
-1.75
-1.8
-1.85
-1.9
-1.95
-2
DNA and Protein Total Energy (kcal/mol)
32
30
28
26
24
22
20
18
16
14
12
10
DNA Curvature (arbitrary units)
sequences in order of increasing number of substitutions

Performance

The prototype model generated and minimized a single configuration with energy calculation in about 305 seconds.

Each protein structure file generation took ~135 seconds, the parallel energy minimization needed ~145 seconds, the energy calculation was ~5 seconds, and the mutation with coordinate preparation and extraneous I/O took ~20 seconds.

If the prototype study were to be done sequentially, the generation and calculation of all 4096 sequences would take 1249280 CPU seconds, roughly 14.5 days.

Performance, cont.

In the parallel study, 20 processors were originally used to operate the framework: one control processor, and 19 worker processors. Since the time a worker spent in energy minimization was about equal to time processing sequential operations, about 10 workers were utilizing four additional processors at a given time running simulations.

Therefore, using 60 total processors, the parallel framework completed the prototype study in 66490 computing seconds, just under18.5 hours.

Performance, cont.

In a parallel run consisting of 19 worker processes, there are 218 rounds of generation necessary to configure all 4096 sequences. Under the same experimental conditions, if all of ITTC's 373 total processors were utilized with 124 worker processes, only 38 rounds of generation would be necessary, and the total time to configure all sequences would be 3.22 hours.

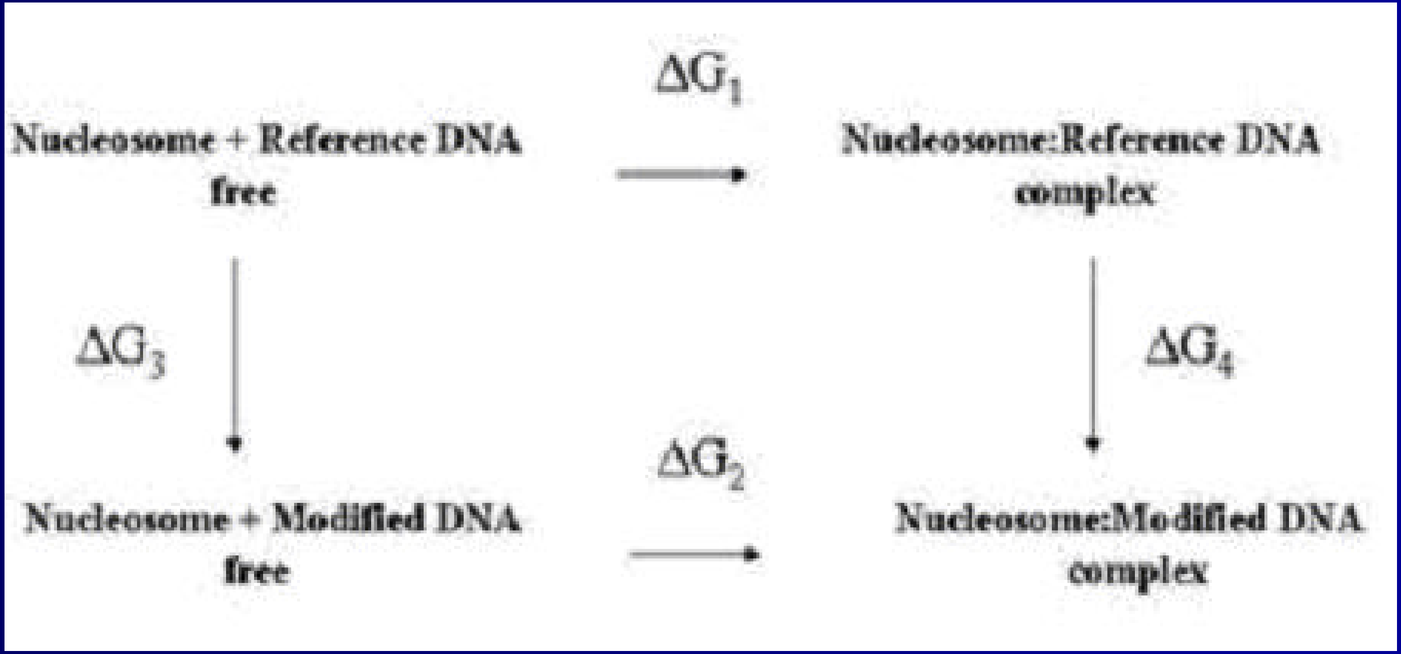
In one month almost a million sequences could undergo configuration and energy calculation about the nucleosome.

Performance, cont.

If the same solvated structure underwent an energy minimization for 6000 steps instead of dynamics, total time for a single configuration would be around ten hours.

With 1160 processors, one could begin analyzing the energies of all 4096 sequences on day 92; with 7565 processors, the time needed to emulate the prototype system would be less than 14 days.

Applications



ΔG_1

Nucleosome + Reference DNA free → Nucleosome:Reference DNA complex

ΔG_3 ↓ ↓ ΔG_4

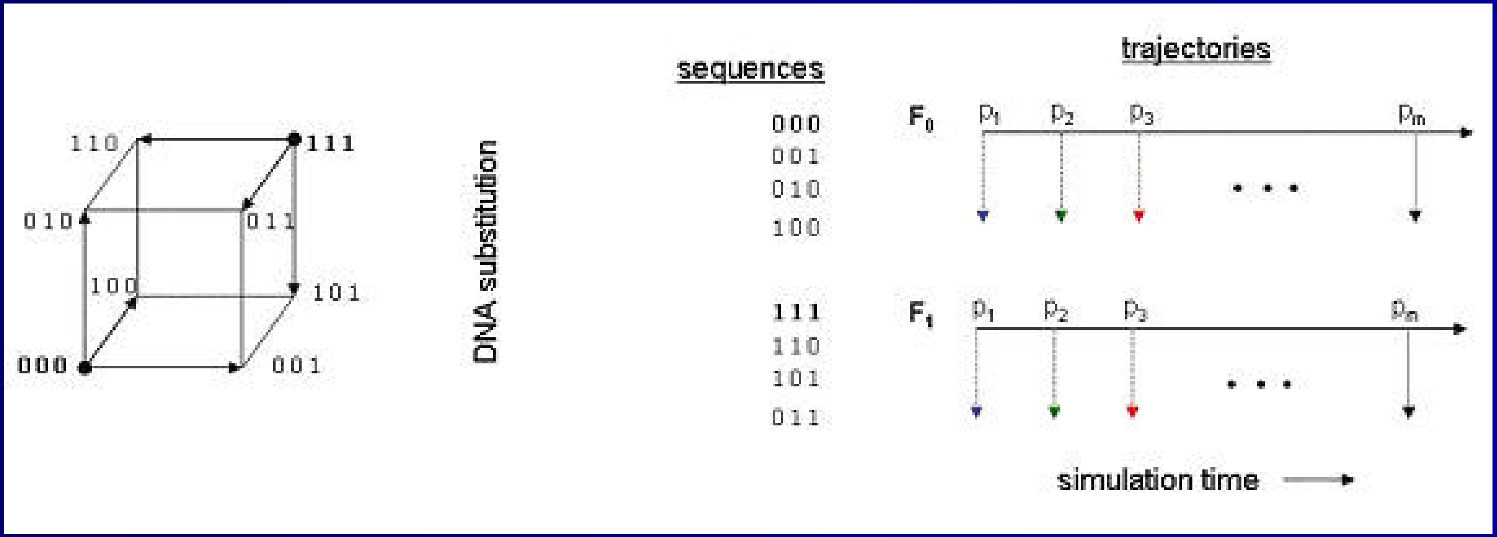
ΔG_2

Nucleosome + Modified DNA free → Nucleosome:Modified DNA complex

A potential study comparing the free energies of a nucleosome complex with that of free DNA and protein could highlight relative binding affinities.

Applications, cont.

Low Perturbation Study. Multiple starting structures can be used in detailed experiments. Notice sequences labeled 000 and 111 form the root node on separate mutation trees. Points along the fully equilibrated trajectories can be used as new starting points to sample the phase space of nucleosome configurations.



Conclusions

Nucleosome research is important to vital cellular events.

The research enables a perspective unto the mechanics of DNA itself.

A framework for the high-performance configuration of many DNA sequences has been built and works.

New and imaginative nucleosome studies can place full attention on experimental design and the analysis of interesting data while the parallel framework configures novel structures automatically.

Thanks

Dr. Terry Clark

Dr. Xue-wen Chen

Dr. Victor Frost

Dr. Krzysztof Kuczera

All ITTC staff and administrators

You!