November 2, 1991

Preliminary report, Laboratory of Molecular genetics and Informatics
The Rockefeller University

Our laboratory research is focussed on ways in which the intracellular environment can influence differential mutagenesis. Our cognitive modelling research is tracking how we approach this problem in day to day laboratory planning, and supports the latter through more systematic plan-generate-test paradigms. These are founded on the DENDRAL style knowledge-based systems. An important element is a hypothesis generator (or classifier) which can expect in principle to generate a tree of all possible hypotheses.

In DENDRAL, this was facilitated by a structure-generating algorithm which could produce all valid molecular structures. In our current effort, we have an n-dimensional phase space where we seek to identify a heuristically useful set of orthogonal variables, each of which can be exhaustively scanned. Later, our plan and test modules will restrict the combinatorial generator to a reasonable scope.

For some time before we implement this in computer programs, we monitor our own laboratory dialectic, seeking the most economical and rigorous axes. Before turning to regionally differential mutagenesis, we have done a logical survey of mutagenesis in general, this being defined as changes in DNA base sequence which are propagated in successive generations of DNA replication in vivo.

After several trials, we concluded that the problem can be dissected into two major branches, each with a number of non-orthogonal components:

a) The attributes of the DNA target
b) The environmental features (biological, chemical, physical) playing on the target.

Under a), we can consider its primary ... quaternary structure (the last being the cellular or extracellular localization). But these components are non-orthogonal, as a given primary sequence can be found in a variety of conformational states.

These attributes also encompass the possible chemical reactions of DNA, and it is helpful to know that DNA structure allows a limited range of chemical alterations (some very familiar, others purely hypothetical). These range from nicking (ss- or ds-) scission of phosphodiester bonds, to reactions of the furanose sugar, to change or extraction of the purine/pyrimidine bases. They can in principle also be followed by religation of the phosphodiester (possibly in configurations other than the canonical 3'-5') and further chemistry of the sugars and bases. At the secondary level, base pairing and ss- to ds- transitions, and v.v., come immediately to mind.

This approach is in its early stages, but we have already found it to be very useful in organizing the enormous mass of relevant information about mutagenesis processes, direct and indirect.

The following lists are a further elaboration of the preceding narrative. Subsequently, we will superimpose a planning perspective that can limit the hypothesis space by heuristics concerning sources of specificity, namely how some DNA (genes, sequences, loci) could be expected to be more impacted than other by an environmental insult or stimulus.

---

1) Characterize the various forms of the substrate (DNA) on which mutagenic processes operate,

2) Identify environmental and cellular signals and reagents which may participate in the process of genetic change,

3) Compile the components of cellular metabolism which act on the substrate, to create a set of primitive transformations which DNA can undergo,

4) Describe mutation-generated changes in primary sequence as the result of the application of these metabolic primitives to the appropriate forms of DNA.

To characterize DNA as a substrate for mutagenesis, it is necessary to identify sets of physical attributes which are orthogonal to each other. Informally, we mean that these physical qualities are mostly independent of each other, and that a given molecule may be characterized by the values of several attributes. Three such important attributes are:

1) The quaternary structure of the DNA target. This includes the strandedness (single, double, triple or quadruple),

2) Conformational aspects such as breathing, writhing, supercoiling status, catenation and cruciform structures. These attributes are the consequence of temperature, local sequence, ionic strength, nucleotide pools, etc.

3) Primary lesions. Examples are strand scission at the phosphodiester bond, deletions, insertions, chemical modifications, and non-covalent intercalations of small molecules.

The features described above (large-scale structure, conformation and characterization of primary lesions) can form the basis for a classification of mutations from the perspective of the substrate. To illustrate, we can look more closely at the third group (primary lesions) and suggest a classification of mutations with respect to to the chemistry of unreplicated DNA (i.e. before its presentation to a polymerase). As we do so, we can anticipate the actions which may result from the presence of such a DNA substrate:

1) Backbone changes -- mutations which interrupt the phosphodiester backbone
   A) Backbone scission of ssDNA (as a lesion in dsDNA)
      Possible actions:

a) religation
b) formation of gap via exonucleolytic action
c) displacement synthesis

B) Backbone scission of dsDNA (can produce blunt or staggered ends)
Possible actions:
a) introduction of a stable end (telomere)
b) recombinogenic repair

2) Changes in which the backbone is not affected -- (base altered instead)
A) Removal of base (depurination)

B) substitution (C -> U, pseudo-bases)

C) base-base union (T dimers)

D) destruction or modification of base

Possible actions:
a) reaction with other ligands
b) interference with replication [modulo recA binding]

Signals and reagents which may influence information metabolism can be classified as follows:

1) Molecules
A) macromolecules
B) small molecules

2) Physical agents
A) emissions through the entire electromagnetic spectrum
B) pressure (including sonic effects resulting from pressure fluctuation)
C) temperature
D) osmolality
E) surfaction
F) pH
G) humidity

3) Physical trauma to cells

4) Electric charge

5) Gravity

6) Surface constraints

We can further describe each of these signals from another perspective, which identifies the spatial, metabolic or physical point of interaction with the cell. Many cellular responses to external signals are mediated by changes in DNA conformation or DNA-protein binding, suggesting a nexus of feedback from environment to differential mutagenesis.

1) Those transduced by a receptor:
   A) membrane
   B) cytoplasm
   C) nucleus -- in bacteria, e.g., enzyme inducers and derepressors

2) Nutrients (C and N metabolism)
   A) elementary sources of C, N, P ...
   B) specific growth factors
   C) light or cognate sources of chemical energy

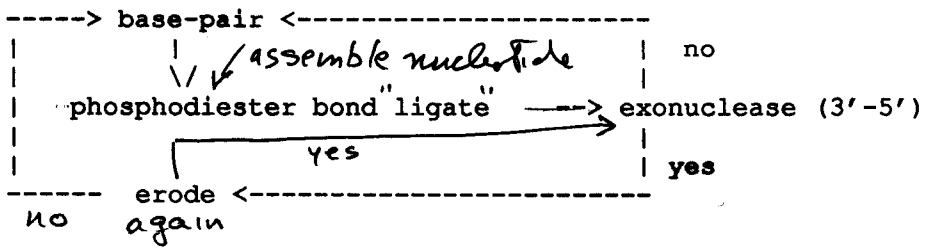
3) Reagents
   A) low molecular weight
      e.g. alkylating agents (mustard gas; formaldehyde; hydrazines)
   B) high molecular weight
      enzymes
      ribozymes
      other nucleic acids (e.g. antisense RNA)

It is then possible to identify action primitives which can alter DNA which has been thus characterized. The cellular metabolic processes which act on DNA include such components as replication, excision repair, nucleolytic erosion, chemical modification, recombination, incorporation of base analogs, and recognition and subsequent repair of mismatch and modifications. We intend to describe these processes as ordered compilation of primitives.

To illustrate primitives, we can examine replication in finer detail. The preconditions on the DNA template include the availability (melting of the helix to form ssDNA), the location of protein binding sites and the presence of a primer. The primitive actions then include base-pairing, ligation and excision.

Given these primitives, it should be possible to describe the transformation of one DNA sequence into another as the result of a series of operations on well-defined states. For example, we can describe the potential sequence changes in a DNA molecule by a state diagram which describes editing. Editing occurs immediately after the addition of a new base to the 3' end of a growing chain. It must happen before ligation of the next base and involves a 3'-5' exonuclease.

Preliminary circuit diagram



This diagram can be elaborated by including endonuclease nicking of DNA, often in response to chemical cross-links, or to base-pair mismatch. The free ends then can be processed just as above.

These finite state machines are a gratifyingly compact summary of systems as diverse as standard DNA-replication (with editting), excisional and mismatch repair, and error-prone repair associated with activation of the "SOS" system.

Most of the detail of this report is taken from Dr. M. Noordewier's notes of weekly discussion sections of the laboratory.

Yours sincerely,

Joshua Lederberg