O que é chado?

* No início, um banco de dados para FlyBase: conjunto de dados incrivelmente interessante
* Um banco de dados para curadoria **muito** profunda
* Um banco de dados integrado
* Um banco de dados que é genérico o suficiente para ser usado por qualquer organismo

Chado - o jeito do chá

Por que a referência ao chá? De acordo com a antiga tradição do GMOD, a lenda diz que os criadores Chris Mungall e Dave Emmert estavam bebendo chá em uma casa de chá quando desenvolveram o primeiro projeto que acabou se tornando Chado.

[](http://gmod.org/wiki/File:Chado.jpg)

Que documentação existe para o chado?

Este Wiki é atualmente a melhor fonte de documentação para Chado. Aqui estão as páginas mais úteis do Wiki:

* [Começando](http://gmod.org/wiki/Chado_-_Getting_Started)
* [Chado Manual](http://gmod.org/wiki/Chado_Manual)
* [Melhores Práticas](http://gmod.org/wiki/Chado_Best_Practices)

Um Esquema Modular

**Lista de Módulos Chado**

* [Auditoria](http://gmod.org/wiki/Chado_Audit_Module) - para trilhas de auditoria de banco de dados
* [Companalysis](http://gmod.org/wiki/Chado_Companalysis_Module) - para dados de análise computacional
* [Contato](http://gmod.org/wiki/Chado_Contact_Module) - para pessoas, grupos e organizações
* [Vocabulário controlado (CV)](http://gmod.org/wiki/Chado_CV_Module) - para vocabulários controlados e ontologias
* [Expressão](http://gmod.org/wiki/Chado_Expression_Module) - para resumos de [expressão](http://gmod.org/wiki/Chado_Expression_Module) de RNA e proteína
* [Geral](http://gmod.org/wiki/Chado_General_Module) - para identificadores
* [Genética](http://gmod.org/wiki/Chado_Genetic_Module) - para dados genéticos e genótipos
* [Biblioteca](http://gmod.org/wiki/Chado_Library_Module) - para descrições de bibliotecas moleculares
* [Mage](http://gmod.org/wiki/Chado_Mage_Module) - para dados de microarray
* [Mapa](http://gmod.org/wiki/Chado_Map_Module) - para mapas sem sequência
* [Diversidade Natural (ND)](http://gmod.org/wiki/Chado_Natural_Diversity_Module) - para vários experimentos, como fenotipagem e genotipagem
* [Organismo](http://gmod.org/wiki/Chado_Organism_Module) - para dados taxonômicos
* [Fenótipo](http://gmod.org/wiki/Chado_Phenotype_Module) - para dados fenotípicos
* [Filogenia](http://gmod.org/wiki/Chado_Phylogeny_Module) - para organismos e árvores filogenéticas
* [Publicação (pub)](http://gmod.org/wiki/Chado_Publication_Module) - para publicações e referências
* [Sequência](http://gmod.org/wiki/Chado_Sequence_Module) - para sequências e recursos de sequência
* [Estoque](http://gmod.org/wiki/Chado_Stock_Module) - para espécimes e coleções biológicas
* [WWW](http://gmod.org/wiki/Chado_WWW_Module) -

Existem *dependências* entre os módulos.

Esta página está focada no [módulo de sequência](http://gmod.org/wiki/Chado_Sequence_Module) ; também discutiremos partes do [módulo cv,](http://gmod.org/wiki/Chado_CV_Module) pois as ontologias são cruciais para como o chado representa todos os dados.

As próprias tabelas chado reais não são discutidas em detalhes de atributo por atributo; isso pode ser navegado verificando o [Manual](http://gmod.org/wiki/Chado_Manual) do [Chado](http://gmod.org/wiki/Chado_Manual" \o "Chado Manual) .

Um dos principais pontos fortes do chado é que ele reúne as visões de mundo da sequência e da genética.

Vamos dar uma olhada na conceitualização do mundo do chado antes de mergulhar no design do esquema.

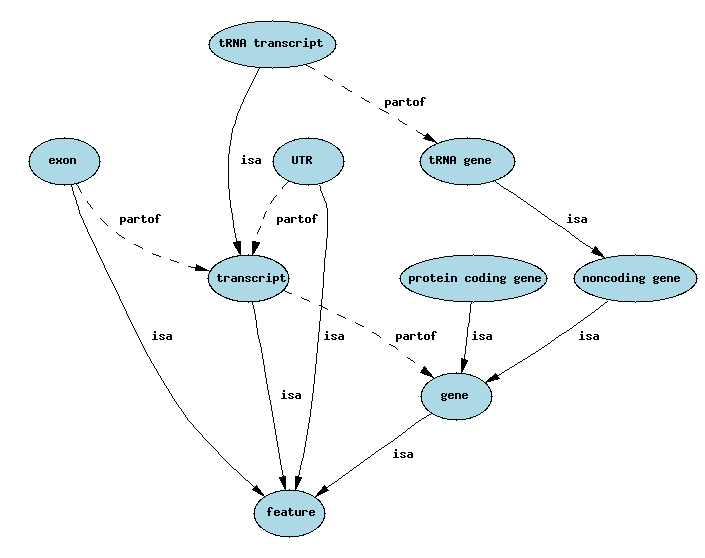
O Módulo de Sequência e Recursos

**Definição**

* Um recurso é uma *coisa*
* Um recurso é *potencialmente localizável*
* Um recurso é ainda definido por uma ontologia

**Tipos de recursos: uma *ontologia***

Uma forma de representar ontologias é por meio de um modelo de grafo, com nós representando conceitos e arestas representando *tipos de relacionamento* entre os conceitos.

**Ontologia de sequência simplificada**  
[](http://gmod.org/wiki/File:So-slim-example.png)

**is\_a** : subtipos, especialização / generalização  
**parte\_of** : composicional

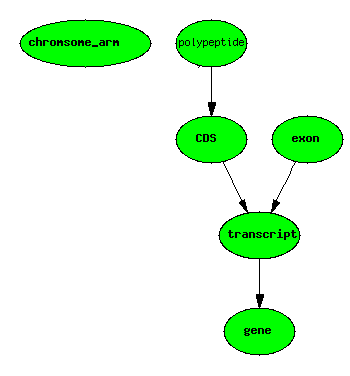
**Alguns outros tipos de recursos**

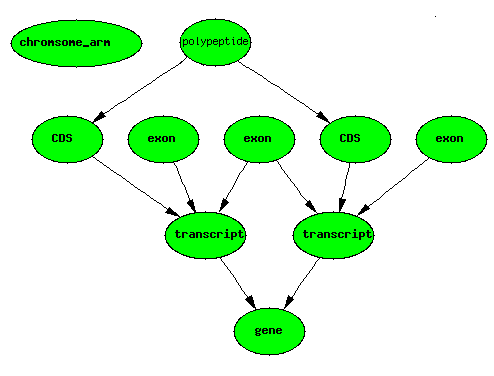
* HSP
* domínio de proteína
* braço de cromossomo
* contig
* andaime
* região reguladora
* características de variação: inserções, exclusões, SNPs

Gráficos de recursos

|  |
| --- |
| **Problema / pergunta** :  As relações composicionais são um subconjunto das relações do gene para o CDS no SO. Como decidir onde simplificar o gráfico? |

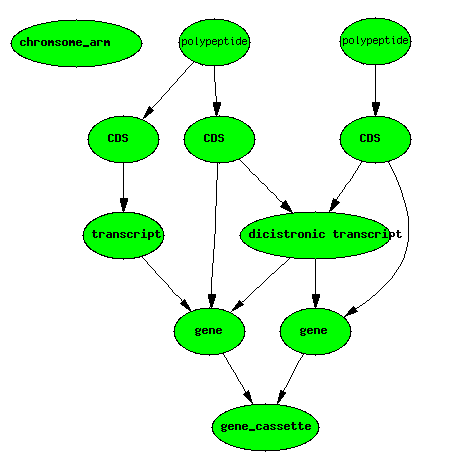
Os nós no gráfico representam instâncias de recursos - os arcos nesses casos representam relacionamentos de composição (embora outros tipos de relacionamento sejam possíveis). Os gráficos de características **não** representam relações posicionais ou espaciais - falaremos disso mais tarde. Para obter mais informações sobre a representação de genes, consulte as [Melhores Práticas de Chado](http://gmod.org/wiki/Chado_Best_Practices) .

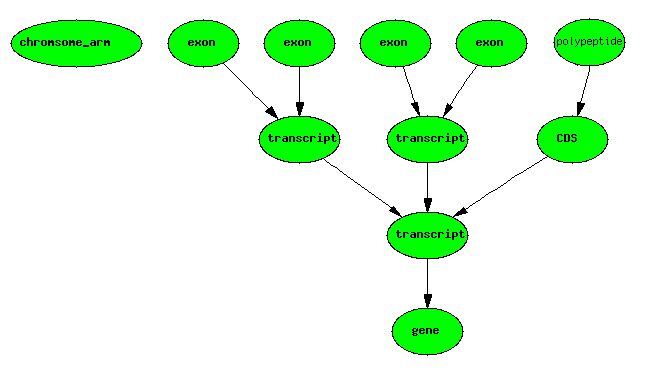
**Exemplo básico de dogma central** Um gene, um transcrito, um exon, uma proteína  
[](http://gmod.org/wiki/File:Basic-central-dogma.png)

**Emenda Alternativa**  
[](http://gmod.org/wiki/File:Alt-splicing.png)

**Gene Dicistrônico**

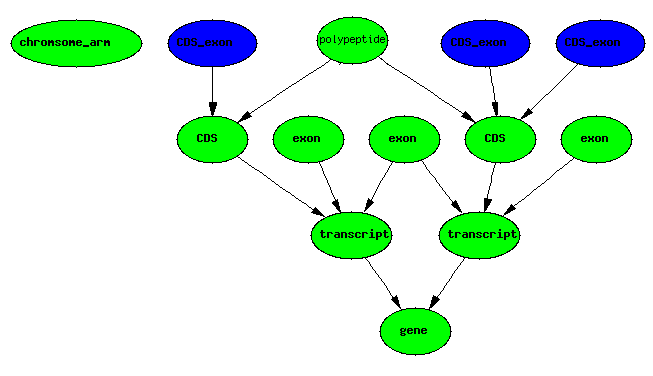
Um gene dicistrônico é um gene com um mRNA que codifica para dois CDSs distintos não sobrepostos. Genes dicistrônicos (ver, por exemplo, os [genes de Drosophila Adh e Adhr <http://www.flybase.org/reports/FBgn0000056.html> ]) têm produtos gênicos totalmente distintos derivados do mesmo transcrito. Para confundir as coisas, os dois polipeptídeos são comumente referidos como derivados de dois genes distintos (por exemplo, Adh e Adhr). Em um banco de dados como [FlyBase <http://flybase.org/> ], existem 3 IDs de genes armazenados no banco de dados - um para cada um dos dois genes não sobrepostos e um para o cassete de gene.

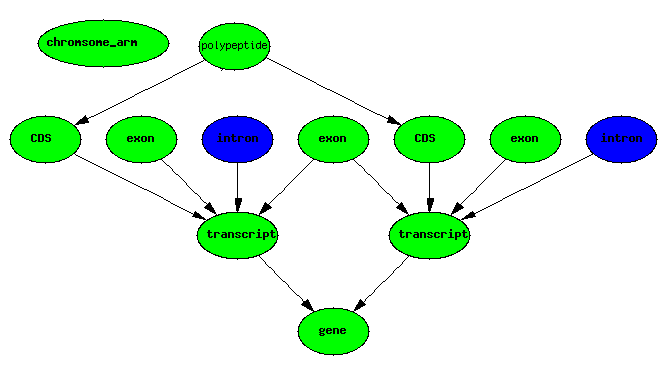
Os genes dicistrônicos tornam difícil ter uma definição formal de gene que corresponda perfeitamente à forma como os biólogos usam o termo.  
[](http://gmod.org/wiki/File:Dicistronic-gene.png)

**Trans-splicing**  
[](http://gmod.org/wiki/File:Transsplicing.png)

Outros casos de trans-splicing podem envolver transcrições primárias distribuídas espacialmente.

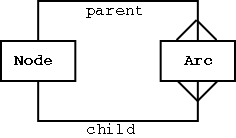
Transformações de gráfico de recursos

**Limites CDS + exões IMPLICA exões CDS**  
[](http://gmod.org/wiki/File:Inferred_cds_exons.png)

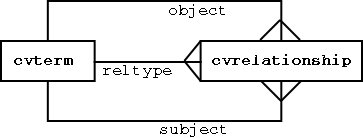
**exons IMPLICA íntrons**  
[](http://gmod.org/wiki/File:Inferred_introns.png)

Representando gráficos em um banco de dados relacional

Um gráfico pode ser definido como uma coleção de **arestas** (arcos) e **vértices** (nós).

**Estrutura de duas tabelas necessária para representar gráficos**  
[](http://gmod.org/wiki/File:Rel-graph.png)

Representando gráficos de ontologia no Chado

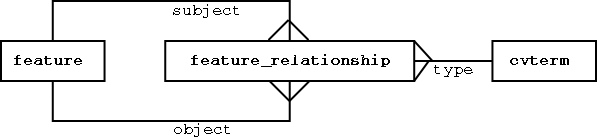
**cvterms (termos de vocabulário controlado) conectados por cvrelationships**  
[](http://gmod.org/wiki/File:Chado-ont.png)

O tipo de relacionamento é um termo controlado em si mesmo. Cada cvrelationship pode ser pensado como uma instrução SUBJECT PREDICATE OBJECT (por exemplo, "GPCR *é-um* transmembrane\_receptor).

A estrutura acima é exatamente a mesma do modelo de dados RDF - muitas linguagens de ontologia modernas (por exemplo, DAML, OWL) são colocadas em camadas sobre RDF, de modo que a estrutura acima garante que seremos capazes de representar todos os conceitos ontológicos mais avançados.

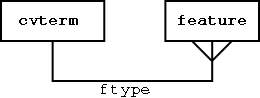
Representando gráficos de recursos no Chado

**features são os nós - feature\_relationships são os arcos**

[](http://gmod.org/wiki/File:Fgraph.png)

Nota: as diferentes classes de recursos podem ser modeladas relacionalmente; o princípio é manter o material estável modelado de forma relacional e o material fluido / extensível modelado em uma ontologia que fica em uma estrutura de banco de dados genérica.

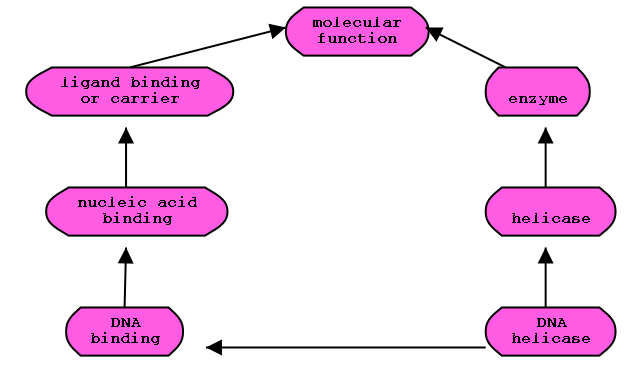
**Os recursos são digitados**

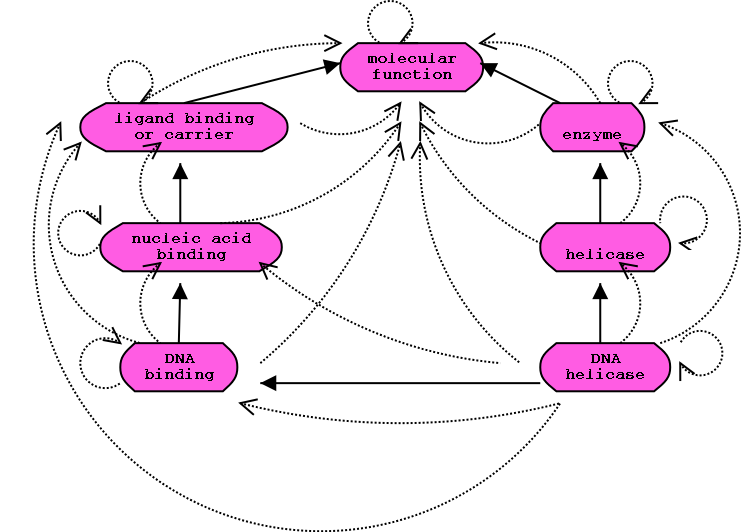
[](http://gmod.org/wiki/File:Chado-feat.png)

Consultando gráficos

A maioria das implementações de SQL *não* é *recursiva* .

Problema: encontre todos os genes; encontre todos os genes (genérico) encontre todos os genes não codificadores encontre todos os genes que codificam proteínas encontre todos os genes de tRNA encontre todos os genes de snRNA encontre todos os genes de snoRNA ... etc eek!

**Solução: pré-computar o**[***fechamento transitivo do***](http://en.wikipedia.org/wiki/Transitive_closure)subgrafo GO Ontology  
[](http://gmod.org/wiki/File:Chado-Go.png)

**Fechamento transitivo do gráfico: as** linhas sólidas representam as relações reais. A coleção de linhas pontilhadas é o fechamento dos relacionamentos.  
[](http://gmod.org/wiki/File:Transitive_closure.png)

forall x SEMPRE TRUE: x R \* x

x R y IMPLICA: x R \* y

x R y, y R \* z IMPLICA: x R \* z

Usando visualizações para simplificar as consultas

A visualização a seguir usa a tabela [cvpath](http://gmod.org/wiki/Chado_Tables" \l "Table:_cvpath" \o "Mesas Chado) , que inclui o encerramento do relacionamento *is\_a* .

**CREATE**  **VIEW** fgene **AS**

**SELECT**

feature . \*

**FROM**

feature **INNER**  **JOIN** cvpath **ON**  ( feature . Ftype\_id = cvpath . Subjterm\_id )

**INNER**  **JOIN** cvterm **ON**  ( cvpath . Objterm\_id = cvterm . Cvterm\_id )

**ONDE** cvterm . termname = 'gene' ;

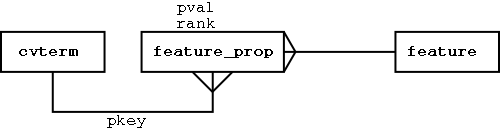
No Flybase, usaremos principalmente chado no modo de *mineração de dados* - isto é, consultaremos, não atualizaremos. Isso significa que podemos *materializar* visualizações para velocidade.

Atributos Extensíveis

O modelo relacional chado define um conjunto fixo de atributos para um [recurso](http://gmod.org/wiki/Chado_Tables#Table:_feature) :

|  |  |  |
| --- | --- | --- |
| **Campo** | **Requerido** | **Descrição** |
| dbxref\_id | não | identificador com espaço de nomes (chave estrangeira) |
| nome | não | Um identificador exclusivo não garantido que é útil para um humano |
| Nome único | sim | nome ou identificador que é único no banco de dados |
| resíduos | não | DNA, RNA ou sequência de proteína |
| md5checksum | não | assinatura de seqüência |
| Seqlen | não | comprimento da sequência (pode estar presente mesmo se os resíduos estiverem ausentes) |
| type\_id | sim | Tipo de recurso de [ontologia de sequência](http://sequenceontology.org/) (chave estrangeira) |

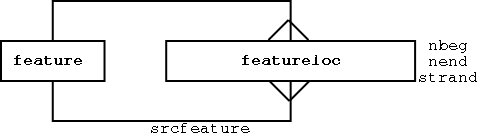
O que acontece se quisermos incluir outros atributos específicos para certos projetos ou específicos para certos tipos de recursos? Podemos usar um paradigma de propriedade de recurso extensível:

**Podemos anexar qualquer propriedade que gostarmos de apresentar:**  
[](http://gmod.org/wiki/File:Fprop.png)

Localização de recursos em coordenadas de sequência

Todas as localizações de sequência são em relação a outro recurso.

Um recurso pode ter vários locais - no entanto, locais "divididos" **não** são usados ​​(para um exemplo de um local dividido, veja como o Genbank representa uma transcrição).

**Qualquer elemento pode ter de 0 a muitos locais:** cada local é relativo a outro elemento (o recurso de **configuração** )  
[](http://gmod.org/wiki/File:Floc.png)

A tabela **featureloc** inclui os seguintes atributos:

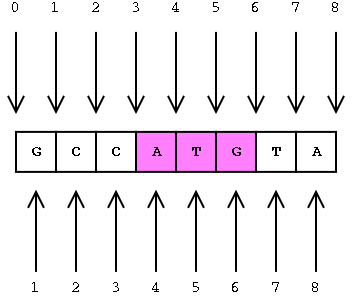
* fmin - limite de 5 'de recursos
* fmax - limite de 3 'de recursos
* strand: a direção do recurso, em relação a srcfeature

as localizações têm *semântica direcional* (como vetores matemáticos). isto é diferente dos min / max semântica utilizados por exemplo [Bioperl](http://bioperl.org/) , GenBank

Coordenadas Interbase

**Chado usa coordenadas *interbase* .**

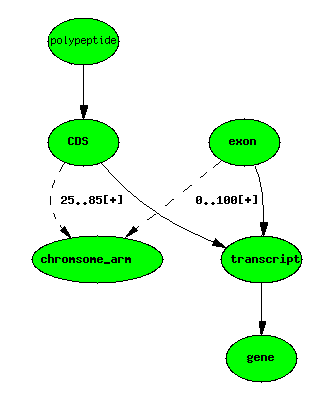
O Interbase conta espaços, não bases. É um sistema de coordenadas baseado em 0; a contagem começa em zero. Este sistema de coordenadas é atraente por uma variedade de razões matemáticas. Os sistemas de coordenadas padrão (por exemplo, banco de genitores, explosão) não permitem uma representação natural de recursos de comprimento zero. Um exemplo de característica de comprimento zero é um local de clivagem que está entre, mas não inclui, 2 aminoácidos. Usando coordenadas interbase, o número total de pares de base em um recurso é calculado subtraindo a coordenada inicial da coordenada final, sem a necessidade de uma adição off-por-1 exigida por outros sistemas de coordenadas. (Em geral, muitas utilidades matemáticas são ativadas pelo número zero.)

**Coordenadas** interbase **(topo) e orientadas para base (abaixo)** A posição do ATG na interbase é [3, 6] (entre a 3ª e 6ª lacunas) A posição do ATG nas coordenadas base é [4, 6] (entre 4ª e 6ª bases inclusive)  
  
[](http://gmod.org/wiki/File:Interbase.png)

Observe as diferentes aritméticas para calcular o comprimento nesses dois sistemas.

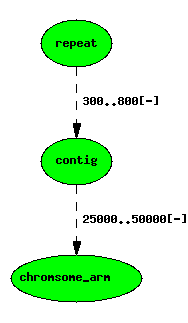
Ao contrário dos vetores matemáticos, devemos também armazenar explicitamente a direcionalidade (fita). Embora isso seja excedente aos requisitos na maioria das vezes, é necessário para recursos de comprimento zero e para cromossomos circulares.

Exemplo básico - com locais

**Dogma central - com exons e CDSs localizados** Usando o princípio de armazenamento mínimo (não armazene nada que não aumente o *conteúdo* de *informações* do banco de dados - ou seja, nada redundante), armazenamos apenas localizações de limites de exon e CDS. Na instanciação do data warehouse do [Berkeley Drosophila Genome Project (BDGP)](http://www.fruitfly.org/) do Chado, podemos escolher armazenar locais para todos os recursos onde forem conhecidos - isso pode simplificar muito algumas consultas, mas deve-se tomar cuidado para garantir que não fiquemos inconsistentes dados.  
[](http://gmod.org/wiki/File:Bcd-loc.png)

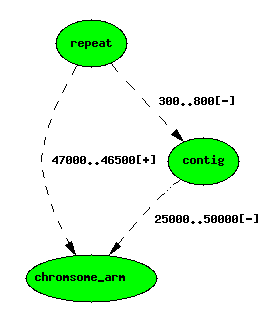
Para a maior parte, inferir os limites dos recursos compostos requer transformações de gráfico bastante simples, embora seja necessário tomar cuidado com os genes que quebram as regras do dogma central.

Os locais podem ser aninhados

**Uma repetição localizada em um contig, ela mesma em um braço do cromossomo, os featureloc** s são representados por linhas tracejadas.  
[](http://gmod.org/wiki/File:Nested-loc.png)

Observe que a posição da repetição no braço do cromossomo é implícita e pode ser calculada com uma transformação de gráfico simples, mas seguindo o princípio de armazenamento mínimo, não armazenamos isso no banco de dados de gerenciamento.

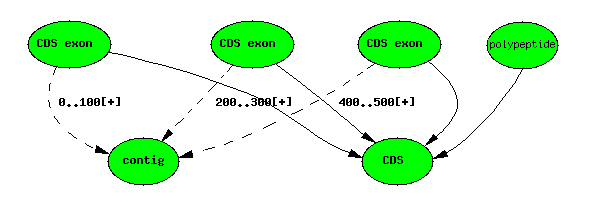
Se desejarmos armazenar a posição redundante em uma cópia para consulta do banco de dados, o chado nos permite esta opção - podemos ter quantos locais quisermos para um recurso. Usamos um atributo extra chamado **locgroup** para distinguir os locais. locgroup = 0 é convencionalmente usado para a localização não redundante.

**O recurso de repetição agora tem dois locais.** Se você olhar para os dados subjacentes, verá que o featureloc que localiza a repetição no braço tem um grupo de valores de **1** .  
[](http://gmod.org/wiki/File:Nested-loc-redundant.png)

Análise computacional: previsões

Todas as previsões são tratadas de forma análoga aos casos padrão do dogma central. Os resultados dos cálculos têm pontuações e são anexados às tuplas na tabela de análise - isso é o que os distingue das "anotações".

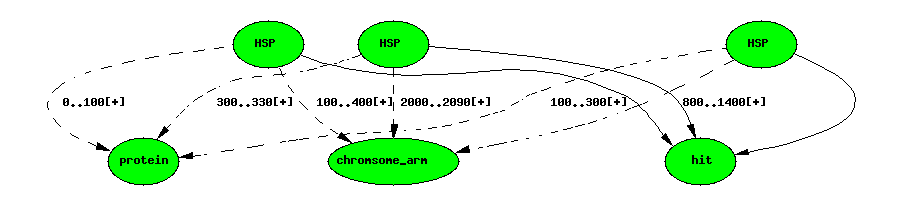
[Genscan](http://www.bioperl.org/wiki/Genscan) prediz CDSs e exões de CDS (não genes no sentido de Ontologia de Sequência). Uma previsão típica do Genscan pode ser assim:

**Predição de 'gene' de 3 exon de Genscan**  
[](http://gmod.org/wiki/File:Genscan.png)

Análise computacional: resultados de similaridade

Similaridades de pares (por exemplo, BLAST) e alinhamentos múltiplos (por exemplo, CLUSTAL) são representados de forma diferente.

Alinhamentos em pares produzem **HSP** s. HSPs são recursos pontuados com dois locais - um na consulta, um no assunto.

**Ataque explosivo com 3 HSPs** Cada HSP tem dois recursos. featureloc tem uma *classificação de* atributo para ordenar os locais; por convenção, 0 é a classificação da consulta e 1 é a classificação do assunto.  
[](http://gmod.org/wiki/File:Blast.png)

Análise computacional: alinhamentos múltiplos

Estes são tratados de forma análoga aos alinhamentos de pares - basta adicionar mais locais.

Características de variação

Os recursos de variação (por exemplo, SNPs, inserções, exclusões) são tratados de maneira semelhante aos alinhamentos de pares.

Mapeamento de bioperl e chado

Existem diferenças importantes entre os modelos Bioperl e chado.

Bioperl permite que um recurso tenha vários locais não contíguos. Mesmo que o esquema chado permita vários locais, anexar vários locais não contíguos é uma violação da semântica chado. Para lidar com isso, criamos recursos extras para as sublocações. Por exemplo, para uma transcrição, criaríamos um exon para cada uma das sublocações.

Se você analisar um arquivo genbank incluindo variações (por exemplo, SNPs) em objetos Bioperl, você obterá um recurso com duas propriedades do tipo "alelo". Isso pode ser representado usando chado; no entanto, a semântica do chado afirma que essas variações devem ser representadas usando vários locais.