Virtual Sequence Database: a constraint logic database for storage of DNA based HLA typing results

Martin Maiers, National Marrow Donor Program, Minneapolis, MN, USA. mmaiers@nmdp.org

Motivation for the development of a Virtual Sequence Database comes from bone marrow registries and is based on the need to 1) maintain accurate and updated DNA based HLA types for volunteer donors who may remain on the registry for many years and 2) select all potential donors who carry HLA alleles that encode MHC molecules which match those of a specific patient. This is challenging because the HLA genes are highly polymorphic and heterozygous. The HLA genes specify molecules with a critical role in tissue compatibility and in defense against pathogens and parasites. These genes are located in the MHC (Major Histocompatibility Complex) which is the most highly polymorphic region of the genome. Because they lack wild-type alleles, MHC genes cannot be described in the simple manner that suffices for monomorphic genes. "Low cost, high volume" PCR-based typing for bone marrow registries targets only subsets of the known sequence polymorphisms. Furthermore new alleles are being continually discovered. A database representation has been developed which addresses these issues by representing the data as a collection of constraints that are defined as statements about the presence or absence of nucleotide subsequences using first order predicate logic. Subsequences are represented as Horn clauses over an alphabet of single nucleotide polymorphisms (SNPs). Cis/trans linkages between polymorphisms are maintained using constraints that represent the heterozygosity of the underlying DNA sequence. The system is designed to store large volumes of data (over 1.5 million samples) collected via several different typing methodlogies (SSOP, SSP, Sequence-based, high-density probe array) and provide a consistent querying interface. This constraint database representation is amenable for searching applications on the basis of sequence features by logical resolution of the target sequence against the data constraints. This system is a precursor to one that can be searched based on features of the secondary protein structure to analyze the effect on the MHC binding repertoire. The virtual sequence representation is also extensible to the tracking of other polymorphic genetic markers elsewhere on the genome and can be seen as addressing the most extreme case of polymorphism faced by the Human Genome Diversity Project (HGDP).

1. Collins, F, et al. New goals for the US Human Genome Project: 1998-2003. *Science*, 1998. 282:682-689.

2. Hansen, JA. Development of registries of HLA-typed volunteer marrow donors. *Tissue Antigens*, 1996. 47:460-463.

3. Helmberg, W, G Lanzer, R Zahn, B Weinmayr, T Wagner, E Albert. Virtual DNA analysis – a new tool for combination and standardised evaluation of SSO, SSP and sequencing-based typing results. 1998. *Tissue Antigens*. 51:587.

4. Hurley, CK. Acquisition and use of DNA-based HLA typing data in bone marrow registries. *Tissue Antigens*, 1997. 49:323-8.

5. Parham, P, T Ohta. Population biology of antigen presentation by MHC Class I molecules. *Science*, 1996. 272:67-74.

6. Robinson, JA. A machine-oriented logic based on the resolution principle. *Journal of the ACM*. 1965. 12:23-41.