A scalable method for efficient stem cell donor HLA genotype match determination

D. Georgiev, L. Houdová, M. Fetter, and P. Jindra

***Abstract*— Finding suitable stem cell donors comprises three independent processes: donor pool HLA typing, donor HLA haplotype inference, and search for donor HLA genotype matches. For practical and technical reasons, these processes are often decoupled leading to informational losses along the way. A method is presented that eliminates some of the technical challenges by considering all three steps together. The method relies on two practical assumptions: there exists a set of common haplotypes and the matched target is typed at high resolution. Under these two assumptions, sufficient statistical HLA analysis of a stem cell donor pool is performed to identify donors that most likely match a given genotype. The presented common haplotype Expected-Maximization (chEM) method is scalable in the number of loci, the number of alleles, and typing ambiguity, overcoming the known curse of dimensionality for the problem of HLA haplotype inference. The practical value is demonstrated on real world data provided by the Czech National Marrow Donor Registry. It is shown the chEM method significantly reduces the field of potential matches when compared to an existing match algorithm.**

***Index Terms*— stem cell donors, HLA refinement, statistical methods**

I. Background

A brief review of the typing, resolution, and matching processes is given. For the sake of simple exposition, unnecessary details are omitted. For more information see [1].

A. Typing methods and nomenclature

Donors are usually typed for up to five loci along chromosome 6: HLA-A/B/C/DRB1/DBQ1. Typing results of donors in a given registry bare different levels of resolution, based on time and location of recruitment. Low level methods rely on antibody-based serological tests. Intermediate resolution molecular methods are based on hybridisation with sequence-specific oligonucleotide probes or PCR amplification with allele specific primers. High resolution methods are based on DNA sequencing tools, which aim to return a four digit number, e.g., A*03:01 or A*03:26, describing the exact amino acid sequence of the corresponding allelic coding region. Ambiguity regarding the amino acid sequence is possible for highest resolution typed donors if their chromosome pair contains an indeterminate allele combination. Decreasing the typing resolution further increases possible allele ambiguity. Hybridisation probes and PCR primers are targeted to a portion of the allele sequence only, and thereby merely identify a list of consistent alleles, e.g., A*03:01-A*30:01. Standard allele lists include entire allele groups, approximately corresponding to broad antigen types, denoted by the first two digits in the allele number, e.g., A*03. Alternatively, specific allele sets are assigned letter codes using the WHO nomenclature [2], e.g., DRB1*15:HMYC. The following is an example typization of a donor from the Czech National Marrow Donor Registry (allele ambiguities, given by the number of possible alleles, are listed in parenthesis):

A	B	C	DRB1
68: JAHD(16)	27: HMHK(13)	07(348)	11: HMXW(12)
03: JAGH(50)	07: HMGA(31)	01(90)	15: HMYC(17)

The above donor is heterozygous at the four typed loci, where the loci HLA-A/B/DRB1 are typed at intermediate resolution and HLA-C is typed at low resolution. The locus HLA-DQB1 is not typed implying it has 509 possible alleles on each chromosome. Such a donor has (at high resolution) 1.9×10^{10} possible haplotypes, 5.3×10^{17} possible HLA genotypes, and 4.3×10^{18} possible haplotype pairs. If the typing data is truncated to low resolution, then the ambiguity is reduced significantly: 224 possible haplotypes, 196 possible genotypes, and 1568 possible haplotype pairs. If, in addition, the haplotype is limited to the typed loci, then (at low resolution) there are 16 possible haplotypes, 1 possible genotype, and 8 possible haplotype pairs.

B. Statistical methods for HLA refinement

Additional typization to increase HLA resolution and reduce the ambiguity (at high resolution) is rarely implemented for a large pool of donors. Instead statistical methods are used to refine a given HLA typing. In the simplest approach, allele frequencies are calculated independently at each locus and used to compute genotype probabilities without accounting for linkage effects. Real world populations, however, share common ancestry and exhibit nonuniform mating patterns that lead to linkage disequilibrium manifested in allele correlations across HLA haplotypes [3].

There is no gold standard for probabilistic haplotype modelling generally accepted by the stem cell donor registries. Three types of methods are available [1]. The Clark's algorithm approaches haplotype modelling parsimoniously beginning with the list of homozygous haplotypes and then

This work was supported by the grant TA ČR TA01010342.

D. Georgiev, L. Houdová and M. Fetter are with the Department of Cybernetics in the Faculty of Applied Sciences, University of West Bohemia in Pilsen, Pilsen 30614, Czech Republic. georgiev@kky.zcu.cz, houdina@kky.zcu.cz, fetter@kky.zcu.cz P. Jindra is with the Biomedical Centre, Faculty of Medicine in Pilsen, Charles University in Prague, Pilsen 30460, Czech Republic, and with HLA laboratory, Czech National Marrow Donors Registry, Pilsen, Czech Republic. jindra@fnplzen.cz

expanding the list in an ad hoc manner until all donor genotypes are resolved. Expected-Maximization (EM) methods consider all possible genotype deconstructions and look for haplotype probabilities of the general population that maximise the likelihood of drawing the donor pool. Bayesian methods rely on more detailed models of population genetics to stochastically generate population haplotypes and empirically estimate haplotype frequencies [4].

Choice of method is determined by several factors. Usually EM and Bayesian methods outperform the Clark's algorithm. Furthermore, EM methods have better convergence properties, include simpler parametrizations, but scale poorly with the number of loci, the number of alleles, and the allele ambiguity. Bayesian methods potentially have lower computational complexity but lack the implementation simplicity of the EM methods. For this reason, most stem cell registries implement their own version of the EM method [5], [6]. Below is a summary of the standard EM algorithm used for haplotype modelling.

The haplotype probability model is iteratively computed as follows.

1) Haplotype set $H = \{h_1, ..., h_n\}$: let d be the number of donors and construct a set of haplotypes by deconstructing the donor genotypes.
2) Initialization of $p_0(k), L_0, s, S, \epsilon$: initialise the probability $p_0(k)$ of haplotype h_k appearing in the population and the likelihood function $L_0 = 0$. Set the iteration count $s = 0$, the maximum iteration count S, and the convergence threshold ϵ.
3) Start: while $s \leq S$, repeat the following steps, otherwise, terminate the algorithm without convergence.
   i) For $k = 1, ..., n$, count the occurrence n_k of haplotype h_k.

$$n_k = \sum_{i=1}^{d} \sum_{j=1}^{n} (1 + \delta_{kj}) P(h_k, h_j | g_i), \qquad (1)$$

   where δ_{kj} is the delta function, $g_i = \{(h_{i11}, h_{i12}), (h_{i21}, h_{i22}), ..., (h_{ir1}, h_{ir2})\}$ is the ith donor's genotype deconstructed into all possible haplotype pairs, and

$$P(h_k, h_j | g_i) = \frac{p_s(k) p_s(j)}{P(g_i)}, \qquad (2)$$

$$P(g_i) = \sum_{j=1}^{r} p_s(h_{ij1}) p_s(h_{ij2}). \qquad (3)$$

   ii) For $k = 1, ..., n$, compute the next iteration of the haplotype probabilities and the likelihood function:

$$p_{s+1}(k) = \frac{n_k}{n}, \qquad (4)$$

$$L_{s+1} = \sum_{i=1}^{d} \log(P(g_i)). \qquad (5)$$

   iii) Check the convergence criteria: if $|L_{s+1} - L_s| > \epsilon$, increment s and return to (i), otherwise, terminate with convergence.

The limitation of the standard EM algorithm is in Step 1. Only a small percentage of donors are typed at high resolution. Deconstructing genotypes of donors typed at intermediate or low resolution into all possible haplotypes generates a prohibitively large haplotype set. Hence, computational complexity attributed to typing ambiguity for actual donor pools is overly prohibitive for simple deployment of the standard EM algorithm [1]. In practice, EM algorithms are either executed at low resolution, where there is little ambiguity (see example above), or executed heuristically to explore specific linkages in partial haplotypes.

C. Donor matching

A donor is a match for a given patient if they share the same genotype. With some exceptions, a patient is typed at high resolution and hence ambiguity arises almost exclusively on the donor side. Two steps are commonly used to resolve this ambiguity. A simple solution is to convert typing data to a lower resolution, e.g., antigen split groups, where there is little ambiguity and common population haplotypes may be used to predict the missing information. Alternatively, the donor genotype is matched only across the typed loci, generating so called 6/6, 8/8, and 10/10 matches. Such matches are of the boolean type, a donor is classified as a potential match if their partially genotype possibly equals the partial genotype of the patient. Effectively, use of probabilistic haplotype models for HLA refinement is ignored.

II. Results

The results in this paper comprise a method that overcomes the greatest shortcoming of the EM algorithm, its computational complexity caused by typing ambiguity at high resolution. Below the common haplotype EM (chEM) method is introduced and its scalability demonstrated in a deployment on a large portion of CNMRD donors typed at various loci and various resolution levels. A five-loci probabilistic model of the Czech population is derived. In addition, the model and the introduced tools are shown to outperform existing intermediate resolution methods in donor matching.

Assumption. The chEM method is based on the following assumptions.

A1) The list of common haplotypes is known.

A2) Patients are typed at high resolution.

Assumption 1 is likely considering most newly discovered alleles are rare [7]. Sequencing methods have also become affordable enough to where Assumption 2 is now generally true.

A. Common haplotype probability model

The derived method differs from the EM algorithm approach described in Section I-B in the following fundamental way. The haplotypes are decomposed into those that are common, denoted simply by $H = (h_1, ..., h_n)$, and those that are rare, denoted by $H^R = \{h_1^R, h_2^R, ...\}$. Rare haplotypes

either contain rare alleles (possibly still unknown) or a rare combination of alleles and as a result have a much lower probability of arising in the population.

The EM algorithm is modified and implemented in the following way to efficiently compute the common haplotype probability model.

1) Haplotype set $H = \{h_1, ..., h_n\}$: the set of common haplotypes is given.
2) Initialization of $p_0(k), L_0, s, S, \epsilon, M$: Initialise p_0 and L_0 as above. Set the values of s, S, and ϵ as above. Set M to be the number of maximum allowable haplotypes per donor.
3) Construct the relevant donor pool $D = \{1, ..., d\}$: a donor in the general pool is deemed relevant as follows.
   i) Set the iteration count $a = 1$ and initialise the donor's list of potential haplotypes H_p to H.
   ii) Consider all alleles possible for the donor at the ath locus and exclude from H_p any haplotype that does not share one of these alleles.
   iii) If the size of $H_p \leq M$, the donor is deemed relevant, otherwise, if a is less than the number of loci, increment a and return to the previous step. If a is equal to the number of loci, the donor is deemed irrelevant.
4) Construct the set of common haplotype pairs $g_i^c, i \in \{1, ..., d\}$: for donor i, first select one of the pre-computed common haplotypes, then search for its pair. The search is done in the same way as when identifying relevant donors (Step 3), the exception being the possible alleles are now limited to the complementary chromosome copies.
5) Continue as in Step 3 of the traditional EM algorithm described above with g^c substituted for g.

Limiting the EM algorithm to the common haplotype list reduces the computational complexity in a number of ways. First, model complexity, i.e., the number of potential haplotypes, is explicitly bounded. As a result, the computational complexity attributed to typing ambiguity (represented by the maximal number of donor haplotypes M) can be independently controlled. In the standard EM algorithm, the number of donor haplotypes grows with M and the number of donors d. Hence to reduce the model complexity, one must consider either high resolution donors or limit the number of donors. Lastly, a relevant donor must not necessarily be typed at high resolution. Their genotypes must merely deconstruct into an acceptable number of common haplotypes. In the case study below, for $M = 150$ the number of relevant donors d from CNMRD is equal to $23,819(54\% of the entire database)$. In the standard EM algorithm, any donor typed at intermediate resolution or below for 2 out of the 5 loci would be deemed irrelevant. In the present, there are only 3,339 (7.6% of the entire database) such donors in the CNMRD. Hence, limiting the model to the common haplotypes greatly increases the statistical relevance of the results for fixed donor ambiguity.

B. Matching genotypes

The chEM algorithm presented in Section II-A only computes the probability model for the common haplotypes. What isn't yet clear is how this model is useful in matching donors to patients, especially when the patient has a rare haplotype. We start by again considering the donor and patient genotypes deconstructed into sets of all possible haplotype pairs g_D and g_P, respectively.

The standard search for a donor with a genotype that likely matches the patient's genotype involves processing the general donor pool and for each donor evaluating the conditional probability

$$P(g_P|g_D) = \frac{P(g_P \cap g_D)}{P(g_D)}, \tag{6}$$

where the Hardy–Weinberg equilibrium is used to compute the probabilities on the right hand side of the equation. For this, however, we need a full probability model of all possible haplotypes in the donor pool. The chEM algorithm of Section II-A only computes the probabilities $P(h_k|\text{donor pool contains 0 rare haplotypes})$, written here simply as $P(h_k|0)$.

If the occurrences of different rare haplotypes in the donor pool are equally likely and statistically independent and ξ is the probability the donor pool has at least one rare haplotype, then the true probability is given by

$$P(h_k) = P(h_k|0)(1-\xi) + \sum_{i=1}^{\infty} P(h_k|i) \left(\frac{\xi}{1+\xi}\right)^i. \tag{7}$$

Under assumptions A1 and A2, the probability ξ is small. Hence we can approximate the match probability in Equation 6 by its Taylor series approximation.

$$P(g_P|g_D) = \sum_{i=0}^{\infty} \frac{\xi^i}{i!} \left.\frac{\partial^i P(g_P|g_D)}{\partial \xi^i}\right|_{\xi=0} \tag{8}$$

Three scenarios are possible and characterised by the dominant terms in the above Taylor series.

Scenario 1) If g_P includes at least one common haplotype pair, then

$$P(g_P|g_D) \approx \frac{\sum_{(h_1,h_2)\in g_P \cap g_D} P(h_1|0)P(h_2|0)}{\sum_{(h_1,h_2)\in g_D} P(h_1|0)P(h_2|0)}.$$

Scenario 2) If g_P includes no common haplotype pairs and at least one pair with one common haplotype, then

$$P(g_P|g_D) \approx \frac{\sum_{(h_1,h_2)\in g_P \cap g_D} P(h_1|0)P(h_2|1)+P(h_1|1)P(h_2|0)}{\sum_{(h_1,h_2)\in g_D} P(h_1|0)P(h_2|0)}.$$

Scenario 3) If g_P includes only rare haplotype pairs, then

$$P(g_P|g_D) \approx \frac{\sum_{(h_1,h_2)\in g_P \cap g_D} P(h_1|1)P(h_2|1)}{\sum_{(h_1,h_2)\in g_D} P(h_1|0)P(h_2|0)}.$$

The above facts suggest that to approximate the donor match probabilities, all rare haplotypes would have to be appended one at a time to the haplotype model. This is true if the probabilities themselves are of interest, however, for the purposes of finding the best donor, only the ordering of the probabilities is required.

Note, for any patient typed at high resolution, it is true that either a donor's possible haplotype pairs include ALL of the patient's possible haplotype pairs or NONE of the patient's haplotype pairs. Correspondingly divide the general donor pool into the potential matches denoted by D_P and the rest, where any donor $i \in D_P$ satisfies $g_i \supseteq g_P$ and any donor $j \notin D_P$ satisfies $g_j \cap g_P = \emptyset$.

Any two donors $i, j \in D_P$ satisfy the equality $g_i \cap g_P = g_j \cap g_p$. As a result, the numerator of the match probability approximations in the three scenarios is constant for all donors in D_P. This leads to the following fact regarding ordering of matching donors.

Fact 1. *Consider the set D_P of donors possibly matching a given patient's genotype, represented by the set of haplotype pairs g_P. As the probability ξ of a rare haplotype in the donor pool approaches 0, for any two donors $i, j \in D_P$,*

$$P(g_P|g_i) > P(g_P|g_j) \text{ if and only if } P(g_i|0) < P(g_j|0), \tag{9}$$

where $P(g_i|0)$ and $P(g_j|0)$ are the genotype probabilities conditioned on there being no rare haplotypes in the donor pool.

The above fact states that only the common haplotype model is required to order the potentially matching donors. The match probability can be estimated using the common haplotype model only when the patient genotype can be reconstructed using common haplotypes. Otherwise, the haplotype model would have to be expanded to include all rare haplotypes.

C. CNMRD derived haplotype model of the Czech population

The haplotype model was derived for a CNMRD donor pool satisfying the ambiguity criterion introduced in the chEM algorithm. CNMRD includes 44,256 donors, of which 39,403 (90%) are typed for HLA-A/B/DRB1 loci and 70% have some serological typization. Donor ambiguity was limited to at most 150 possible common haplotypes, met by 54% of the registry's donors. Donors deemed overly ambiguous to yield significant statistical information included many that were typed for only two loci, e.g., HLA-A/B, and had up to 20 thousand possible common haplotypes.

The list of common haplotypes making up the derived haplotype model included $64,800$ entries taken from [8]. This was assumed to be the list of all common haplotypes called for by Assumption 1. In deployment, however, the possible genotypes of numerous donors from the selected donor pool could not be reconstructed using the listed common haplotypes. All such donors had serologically typed HLA-C. Consultation with field experts revealed that older serological typing methods for HLA-C have a 30% error rate. Once HLA-C typing was ignored for these donors, common haplotype pairs were successfully used to reconstruct at least a single genotype for every donor in the selected pool.

A prototypical algorithmic implementation was built using the open source PosgreSQL database system. The convergence criterion ϵ was set to $3.8 \times 10^{-4}\%$ of the final log-likelihood value. Moreover, a threshold, below which a probability was taken to equal zero, was set to 1×10^{-8}. A naive implementation of the chEM algorithm converged in 22 iterations in approximately 36 hours. While this execution time seems excessive, the common haplotype model does not need to be re-computed in real time. In addition, much faster convergence is expected when performing only model updates.

The final computed model included 11,458 haplotypes with positive probability. The following is a list of the five most common haplotypes and their probabilities identified for the Czech population.

A	B	C	DRB1	DQB1	Prob.
01:01	08:01	07:01	03:01	02:01	0.050
03:01	07:02	07:02	15:01	06:02	0.012
02:01	13:02	06:02	07:01	02:01	0.010
23:01	44:03	04:01	07:01	02:01	0.009
01:01	57:01	06:02	07:01	03:03	0.009

D. Validation

The method was tested on real patient case studies and the matching results were compared against outputs of existing intermediate level search tools. In searching for matches, a larger donor pool was used. Any donor typed for at least HLA-A/B/DRB1 was considered in the search (this included 90% of the donors contained in the CNRMD database). The ignored database entries include 1,595 donors typed serologically only for HLA-A/B and 3,025 donors typed for HLA-A/B/C, where often the HLA-C typing is erroneous. In practice, the minimal match criterion used is 5 out of 6 matching alleles. Hence, the ignored donors are most likely not suitable candidates. Of the donors that were included in the search, 3,168 were typed at high resolution, 22,848 were typed for HLA-A/B/C/DRB1, and 13,299 were typed for HLA-A/B/DRB1.

The search method currently in use was implemented over 15 years ago to find matching donors using their serological typing data. The method relies on uniform typing data in terms of resolution and HLA loci. Therefore, typing of donors obtained using molecular methods is converted to serological format using standard conversion tables and genotypes are only matched across HLA-A/B/DRB1 loci. A donor with potentially matching alleles at all three loci and both chromosomal copies is called a 6/6 match.

The chEM method presented herein searches for matches across all loci, doesn't require uniform typing data, and is capable of finding high-resolution matches. For patients with rare genotypes, the method simply ranks the potential donors. For patients with common genotypes, the method ranks the donors and lists their match probabilities. Nonetheless, for comparison purposes, the method was naturally reduced to search for matches across the HLA-A/B/DRB1 loci.

The CNMRD database 6/6 matches identified by the two methods were compared for 100 patients chosen at random. The existing method relies on serological conversion. Hence it is expected that the presented chEM method will shrink the set of potential donors. The table below summarises this comparison study.

#	RESULT
41	cases where matches were found using existing method
29	cases where matches were found using the chEM method
21	cases where chEM decreased the number of matching donors

As expected, in the majority of cases, the chEM method reduced the set of matching donors. In no cases, the chEM method identified matching donors not identified by the existing method. Out of the cases where the chEM method decreased the number of potential matches, the average reduction was approximately 25.8%.

Below is an example patient and the most likely 6/6 matching donor identified by the chEM method. Despite the donor being typed only at intermediate resolution (the donor has 2×10^{20} possible genotypes) for the three loci HLA-A/B/DRB1, the probability of a 6/6 match is equal to 0.97. Using the chEM method, a search for 10/10 matches was also performed. The same donor was also the most likely 10/10 match with match probability equal to 0.11.

	A	B	C	DRB1	DQB1
Patient	02:01	27:05	01:02	01:01	05:01
	02:01	38:01	12:03	13:01	06:03
Donor	02: KDMZ	27: KCXM		13: KEGB	
	02: KDMZ	38: KCXZ		01: KEFC	

III. Discussion

One of the fundamental limitations of statistical HLA typing resolution refinement is the computational complexity. Typing ambiguity present in donor pools is almost always prohibitive forcing the use of heuristic reductions. Herein we demonstrate that haplotype models used for refinement need to be only defined over the common haplotypes. As long as the probability of rare haplotypes is sufficiently small, potential donor matches can be identified and sorted purely on the bases of the common haplotype model. A prototypical implementation of the method was proven in terms of robustness and computational speed for the CNMRD database of over 40 thousand donors. The method was further validated in matching donor search case studies, where the results produced using existing donor search tools were significantly refined in the majority of cases.

References

[1] J. Listgarten, Z. Brumme, C. Kadie, G. Xiaojiang, B. Walker, M. Carrington, P. Goulder, and D. Heckerman, "Statistical resolution of ambiguous HLA typing data," *PLoS Comp. Biol.*, 2008.

[2] *Nomenclature for Factors of the HL-A System*, WHO nomeclature committee, 1968.

[3] A. Degioanni, P. Darlu, and C. Raffoux, "Analysis of the french national registry of unrelated bone marrow donors, using surnames as a tool for improving geographical localisation of HLA haplotypes." *European Journal of Human Genetics*, 2003.

[4] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, 2001.

[5] D. Steiner, "Probabilistic matching in search for unrelated hematopoietic probabilistic matching in search for unrelated hematopoietic stem cell donors," Ph.D. dissertation, Czech Technical University in Prague, 2013.

[6] Z. S. Qin, T. Niu, and J. S. Liu, "Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms," *American Journal of Human Genetics*, 2002.

[7] P. Jindra, P. Venigová, L. Houdová, and K. Steinerová, "A novel HLA-A null allele (A*02:395N) with stop codon in exon 2 generated by single nucleotide exchange," *Tissue Antigens*, 2013.

[8] L. Gragert, A. Madbouly, J. Freeman, and M. Maiers, "Six-locus high resolution hla haplotype frequencies derived from mixed-resolution dna typing for the entire us donor registry," *Human Immunology*, 2013.