Title

A CASE-BASED REASONING SYSTEM FOR GENOTYPIC PREDICTION OF HIV-1 CO-RECEPTOR TROPISM

Journal

Journal of Bioinformatics and Computational Biology, 11(04)

ISSN

0219-7200

Authors

EVANS, MARK C
PAQUET, AGNES C
HUANG, WEI
et al.

Publication Date

2013-08-01

DOI

10.1142/s0219720013500066

Peer reviewed

A CASE-BASED REASONING SYSTEM FOR GENOTYPIC PREDICTION OF HIV-1 CO-RECEPTOR TROPISM

Mark C. Evans[1], Agnes Paquet[1], Wei Huang[1], Laura Napolitano[1], Arne Frantzell[1], Jonathan Toma[1], Eric Stawiski[1], Matthew B. Goetz[2], Christos Petropoulos[1], Jeannette Whitcomb[1], Eoin Coakley[1], Mojgan Haddad[1,*]
1: Monogram Biosciences Inc., South San Francisco, CA 94080, USA
2: VA Greater Los Angeles Healthcare System and David Geffen School. of Med. at UCLA, CA, USA

*Corresponding author:
Mojgan Haddad
Monogram Biosciences
345 Oyster Point Blvd., South San Francisco, CA 94080
Phone: (650) 616-3645
Fax: (650) 616-3652
Email: mhaddad@monogrambio.com

Keywords: Tropism prediction; Genotypic algorithm; HIV-1 co-receptor; V3.

Running Head: Case-Based Reasoning for HIV Tropism Prediction

Abstract

Accurate co-receptor tropism (CRT) determination is critical for making treatment decisions in HIV management. We created a genotypic tropism prediction tool by utilizing the case-based reasoning (CBR) technique that attempts to solve new problems through applying the solution from similar past problems. V3 loop sequences from 732 clinical samples with diverse characteristics were used to build a case library. Additional sequence and molecular properties of the V3 loop were examined and used for similarity assessment. A similarity metric was defined based on each attribute's frequency in the CXCR4-using viruses. We implemented three other genotype-based tropism predictors, support vector machines (SVM), position specific scoring matrices (PSSM), and the 11/25 rule, and evaluated their performance as the ability to predict CRT compared to Monogram's enhanced sensitivity Trofile® assay (ESTA). Overall concordance of the CBR based tropism prediction algorithm was 81%, as compared to ESTA. Sensitivity to detect CXCR4 usage was 90% and specificity was at 73%. In comparison, sensitivity of the SVM, PSSM, and the 11/25 rule were 85%, 81% and 36% respectively while achieving a specificity of 90% by SVM, 75% by PSSM, and 97% by the 11/25 rule. When we evaluated these predictors in an unseen dataset, higher sensitivity was achieved by the CBR algorithm (87%), compared to SVM (82%), PSSM (76%), and the 11/25 rule (33%), while maintaining similar level of specificity. Overall this study suggests that CBR can be utilized as a genotypic tropism prediction tool, and can achieve improved performance in independent datasets compared to model or rule based methods.

Introduction

Human Immunodeficiency Virus type 1 (HIV-1) gains entry into the human host cell by using CXCR4 (X4) or CCR5 (R5) co-receptors [1]. Given the availability of CCR5 antagonists as a treatment option [2], it is critical to have diagnostic assays available that quickly and accurately determine the co-receptor tropism in a clinical setting. Several studies have been conducted to identify the genetic basis for virus' preference in co-receptor usage, and narrowed down the primary determinant of tropism to the 35 amino-acid of the third hypervariable (V3) loop of HIV-1 envelope [3]. Genotype based prediction of virus tropism utilizing the sequence of the V3 loop offers a rapid test for co-receptor usage. To date, many bioinformatics methods for tropism prediction have been developed and tested. These bioinformatics predictors include support vector machines (SVM) [4, 5], neural networks (NN) [6], decision trees [7], random forest [8], instance based reasoning [9], position specific scoring matrices (PSSM) [10], multiple linear regression [11], and the 11/25 rule [12]. However, these methods generally are developed by fitting a model onto the respective training set, and might not perform as well in independent or unseen datasets [13]. Moreover, as previously reported [14], some of these methods were trained on clonal sequences, and may not be adequate for tropism testing in clinical isolates that are often heterogeneous and have high levels of sequence ambiguity.

In this study, we developed a novel bioinformatics algorithm for genotypic tropism prediction utilizing the case-based reasoning (CBR) technique. CBR [15, 16] originated in the early eighties and was quickly adopted into a wide range of disciplines, from solving routine resource disputes as implemented in MEDIATOR [17] to assisting with medical diagnosis [18]. A case-based reasoner attempts to solve new problems with an unknown solution by adapting established solutions to similar problems. CBR appears to be particularly promising as a genotype based tropism prediction method, as it directly utilizes the genotypic information from clinical specimens generated thru bulk or clonal sequencing, without extrapolating a model or rule set from the data. The high dimensionality of the genetic space as well as the complexity of the co-receptor usage pose a challenge for inferring a good mathematical fit or a set of rules to explain the tropism. CBR operates as a heuristic process that performs guided retrieval and utilization of prior experiences, particularly, pairs of a V3 sequence with the phenotypic CRT assessment to perform *in-silico* tropism prediction.

Methods

Case-Based Reasoning

Case-Based Reasoning (CBR) is an artificial intelligence technique that solves new problems based on the solutions to similar past problems. Following steps were performed to build a CBR algorithm for genotypic tropism prediction: 1) a case library of HIV-1 specimens was compiled from which V3 sequences and phenotypic tropism assessments were obtained; 2) the input problem was characterized by identifying amino-acid sites and physiochemical characteristics of the V3 sequence highly associated with tropism, and weights were assigned to each selected feature for use in similarity assessment; 3) finally, a process for retrieving relevant cases from the library and generating a tropism prediction according to the most similar cases was implemented. These steps are outlined in detail in the following sections.

Data Collection and Construction of the Case Library

All V3 sequences available in Monogram's database were obtained; these include samples from the patient testing database for which genotypic data was available, as well as a cohort of treatment experienced patients (TORO), plus a treatment naïve cohort (LTM) [19]. In order to eliminate potential influences from data variability, sequences and tropism assessments from a central lab (Monogram) with consistent quality were used in all analyses performed in our study. V3 sequences were derived using population sequencing in Monogram's research lab. In the case of amino-acid mixtures, the ambiguity was resolved in favor of the amino-acid more prevalent in the X4 tropic set using a PSSM model we developed based on the previously described method [10]. Amino acid insertions and deletions were coded with an insertion or a gap character (Z and -), respectively. All final sequences were of length 35 to allow comparison of amino-acid sequences position by position. Duplicate sequences were removed from this set. Phenotypic co-receptor tropism was

determined by the Monogram Biosciences' Enhanced Sensitivity Trofile® Assay (ESTA) [20]. In all, 1012 unique V3 sequences were identified from as many patients, resulting in 595 R5 tropic and 417 Dual/Mixed (DM) or X4 tropic viruses. Out of these, 732 (406 R5, and 326 X4/DM) predominantly subtype B samples were selected for training purposes. Two sets of samples were set aside for testing, both with unknown treatment history: a set of 152 commercial samples of mostly subtype B, and one set of 128 clinical specimens with subtype C.

Subsequently, we examined additional sequence characteristics, such as the count of nucleotide and amino-acid mixtures in the original sequence and peptide statistics, and performed statistical analysis to evaluate the importance of these attributes relative to tropism determination.

Given that the co-receptors CCR5 and CXCR4 are different proteins with different physiochemical characteristics in the local environment of their V3 binding sites, we were interested to explore whether there were significant physiochemical shifts in the nature of the V3 peptide that correlated with co-receptor usage. Physiochemical properties of the V3 amino-acid peptides were determined using the Pepstat program [21].

Profile Hidden Markov Models (pHMM) are statistical models of multiple sequence alignments. It was of particular interest to isolate the R5 specific characteristics (or "R5-ness") of the sequence in a measurement. We therefore used treatment naïve, R5-using samples to minimize the possible impurity of the virus population resulting from treatment exposure. Using this subset of samples, a multiple sequence alignment was created and was used to generate a pHMM by applying the HMMR 3.0 application suite [22-24].

We then examined these additional attributes using univariate analysis to identify features significantly associated with co-receptor usage.

Univariate Analysis for Feature Selection

In order to identify significant associations between a given attribute and co-receptor usage, Fisher's Exact Test (FET) was performed and an odds ratio was calculated based on presence or absence of a feature in the X4-using set. Mutations and attributes with strong association with tropism, as identified by FET, were included as the data fields of the case library. We used the log of the odds ratio to assign a weight to every position in the amino-acid sequence as well as all selected features.

Similarity Metric and Adaptation

To evaluate a query sequence against the case library, the query was compared to each member of the case library. All amino-acid positions were examined for a match between the query sequence and the case in the library. For attributes that describe the sequence characteristics as a continuous value, similarity was defined as a range for the absolute difference. Based on the log of the odds ratios calculated in the FET analysis, we generated an array of weights for the 35 amino-acid sites as well as the additional features in the case library. When performing the comparison between a new problem and the cases stored in the library, for each identical amino-acid and for every similar feature, the respective weight was added to calculate a total similarity score.

The adaptation strategy was fine-tuned to maximize the X4 sensitivity. Based on the similarity scores calculated for all cases in the library, if any of the top three scoring cases is DM or X4, then the query is predicted to be X4-using, otherwise it is called R5-using.

Evaluation of the Method

We obtained performance characteristics for the CBR system using a leave-one-out (LOO) approach, excluding the query sequence from the case library and executing the test on the remaining cases in the library. Accuracy of the CBR system was evaluated as the ability to predict CRT compared to ESTA. We used 2 independent datasets to further evaluate the algorithm in unseen data: one comprised of 152 mostly subtype B, and one set of 128 subtype C samples.

In addition, we compared the performance of CBR with SVM, PSSM, and the 11/25 rule [12], to include previously utilized bioinformatics methods with a range of reported performance characteristics. To allow a fair comparison between the methods, we used the same training set as the CBR system to generate PSSM and SVM models, rather than utilizing available methods such as geno2pheno with their existing models [5].

To construct the SVM, the V3 sequences were coded into a vector of length 35 x 22 containing 0 or 1 at each position to describe the amino-acid composition. Counts of selected nucleotide and amino-acid ambiguities were used as additional input parameters. The SVM model was trained using libsvm in R package e1071 (linear kernel; cost=0.35). The cutoff for SVM decision values was optimized through ROC analysis [25]. The PSSM model was developed according to the previously published method [10]. We also applied the 11/25 rule [12], which is based on the presence of amino-acids K or R at position 11, and R at position 25, on our datasets.

Results

A case-based reasoning (CBR) system was constructed to perform tropism prediction based on the V3 loop of the HIV-1 sequence. Our CBR algorithm consists of a case library of 732 V3 sequences with a matched phenotype as determined by ESTA. We first extracted sequence characteristics that provide additional information about the co-receptor usage. The following groups of sequence features were evaluated: 1) count of nucleotide and amino-acid mixtures in the V3 loop; 2) peptide statistics; 3) score generated from the pHMM developed based on a set of treatment-naïve and R5-tropic samples.

Sequence and Physiochemical Characteristics

The correlation between sequence length as well as amino-acid and nucleotide mixtures and tropism was evaluated using univariate analysis. As displayed in Figure 1A, a strong association was found between DM tropism and presence of the nucleotide ambiguities R, Y, W, and K, as well as mixed amino-acids (X).

The Pepstats program was used to analyze each sequence in the case library and to determine its physiochemical profile. Among the characteristics calculated were molecular weight, net charge (charge) and isoelectric point (iep). Additionally the molar composition by biochemical class (aliphatic, aromatic, polar/non-polar, charged, basic, acidic, tiny and small) of the V3 peptide was evaluated. The distribution of each characteristic across the three tropism groups is shown in Figure 1B. The graphs for net charge and iep show similar profiles, as does the charged amino-acid group. This is expected since the isoelectric point is driven by the net charge, which is in turn driven by the percentage of charged residues that comprise a peptide. Inferences about the nature of the charged residues are made by comparing the basic and acidic composition graphs, with the graph for basic residues resembling the pattern for charged, charge and iep. This suggests a preference for X4/DM tropic viruses to have V3 sequences that are more basic in nature and to have a more basic local isoelectric point. Among the peptide statistics, charge, basic, iep, and small groupings showed strong distinction between X4/DM and R5 tropisms (odds ratios=5.1, 5.0, 4.1, and 0.3 respectively; Bonferroni corrected p-values < 0.001).

In an attempt to capture the “R5-ness” of the virus, we obtained and examined a score generated based on a pHMM that we developed using a subset of treatment naïve and R5-using samples. This score is referred to here as the HMM Score. As shown in Figure 1C, pure X4-tropic viruses have distinctly lower HMM scores compared to the R5 and DM sets, while scores derived from DM viruses are generally lower than R5-using samples, but higher than the X4-tropic set.

Amino acid positions and substitutions as well as quantity and quality of sequence ambiguities were evaluated using Fisher’s Exact test. Graphical representation of the features significantly (Bonferroni adjusted p-value<0.05) associated with co-receptor tropism and their weight, as derived from the odds ratios, is shown in Figure 2.

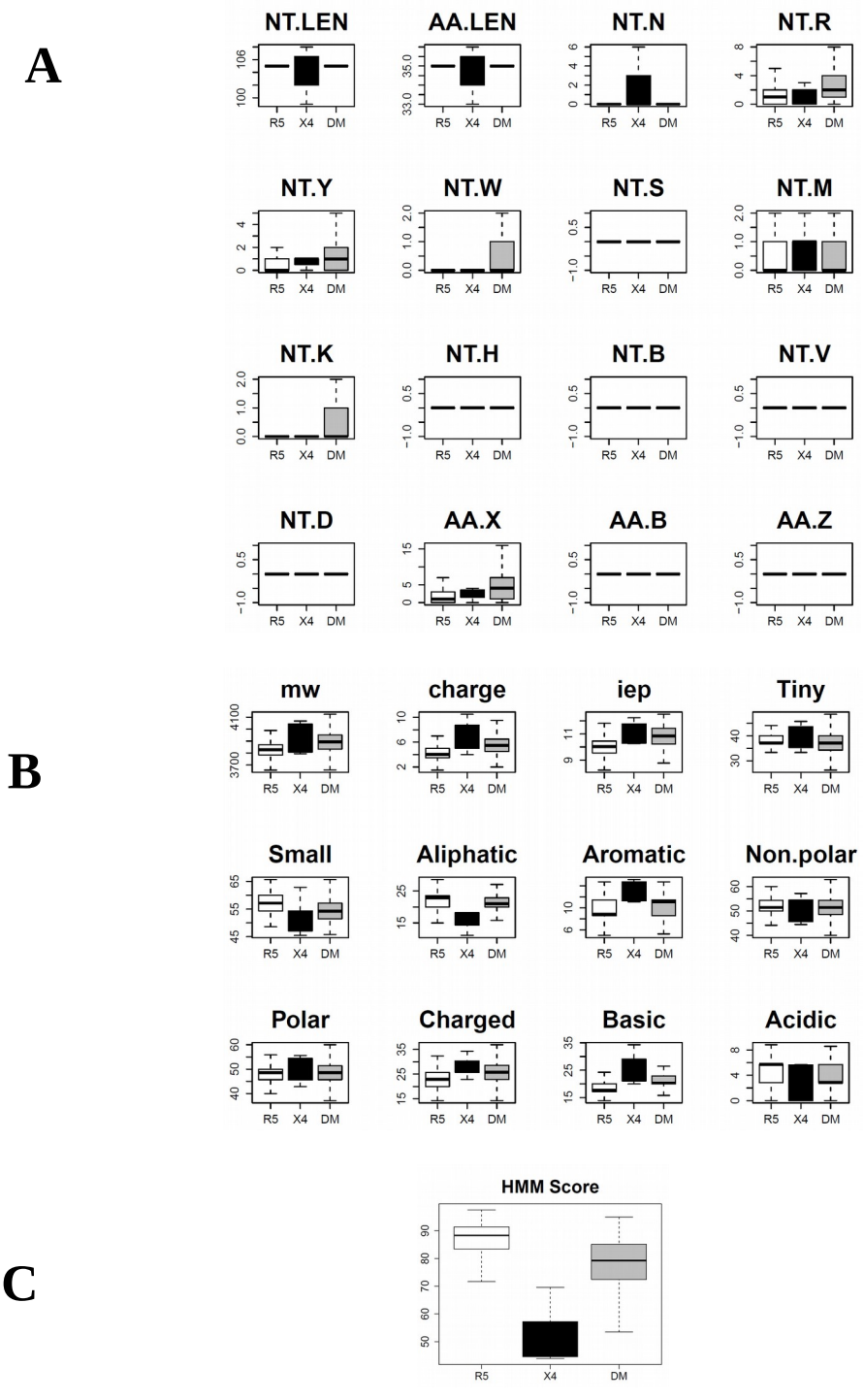

Figure 1: Physiochemical and sequence characteristics: **A)** Sequence length and count of nucleotide and amino-acid ambiguities grouped by tropism as determined by ESTA. Abbreviations NT = nucleotide ambiguity, AA = amino-acid ambiguity. Len refers to sequence length. All other letters correspond to IUPAC ambiguity codes. **B)** Distribution of physiochemical characteristics grouped by tropism as determined by ESTA. mw= molecular weight, charge = Net charge, iep = Isoelectric point, Tiny = (A,C,G,S,T), Small = (A,B,C,D,G,N,P,S,T,V), Aliphatic = (A,I,L,V), Aromatic = (F,H,W,Y), Nonpolar = (A,C,F,G,I,L,M,P,V,W,Y), Polar = (D,E,H,K,N,Q,R,S,T,Z), Charged = (B,D,E,H,K,R,Z), Basic = (H,K,R), Acidic = (B,D,E,Z). **C)** Distribution of R5 HMM score in the case library grouped by tropism.

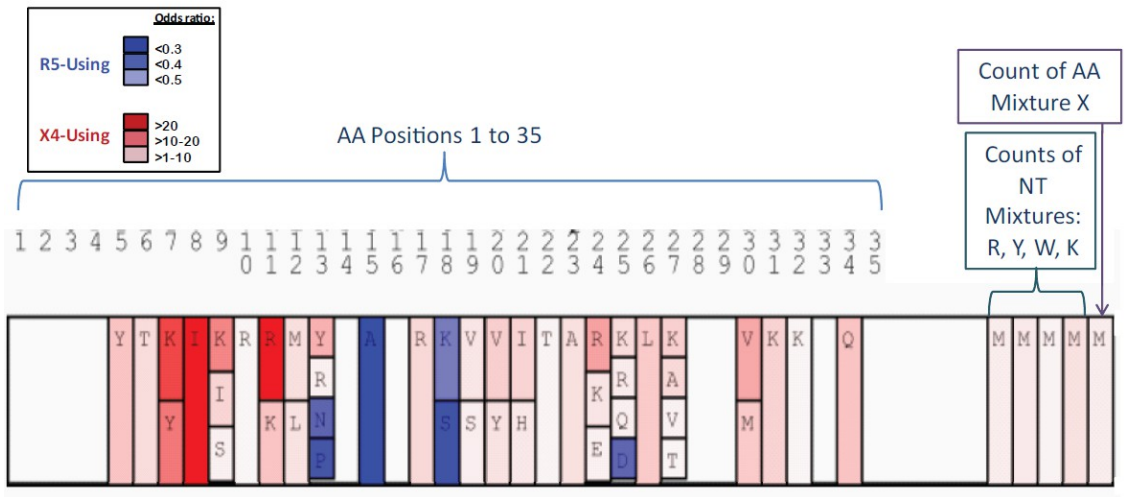

Figure 2: Selected features and weights based on Fisher's Exact test results. Red to light pink = high to low association with X4-usage, Blue to light-blue = high to low association with R5-usage.

Performance

The predictive power of the CBR system was evaluated by performing LOO process in which each of the 732 cases was temporarily removed from the case library and was then presented to the CBR tool as a case for tropism prediction. We obtained true positive (TP) rate or sensitivity, true negative (TN) rate or specificity, number of false positive (FP) and false negative (FN) cases, and the overall concordance as compared to ESTA. These results are shown in Table 1A. The CBR algorithm achieved a specificity of 73.2%, a sensitivity of 89.6% and an overall accuracy of 80.5% in the training set. Furthermore, the training set of 732 V3 sequences was used to construct a PSSM, and a SVM model. The 11/25 rule was also applied to this dataset. For comparison, the performance of these additional predictors is shown in Table 1A. The sensitivity of the CBR algorithm was the highest among the bioinformatics predictors we investigated.

Table 1A: Comparison of performance in the training set

	Prediction Method	TN	FP	FN	TP	Concordance	Spec.	Sens.
Training set (N=732)	CBR	297	109	34	292	80.5%	73.2%	89.6%
	SVM	366	40	48	278	87.9%	90.1%	85.3%
	PSSM	305	101	63	263	77.6%	75.1%	80.7%
	11/25 Rule	392	14	208	118	69.7%	96.5%	36.2%

Table 1B: Comparison of performance in an unseen dataset

Unseen set	Prediction Method	TN	FP	FN	TP	Concordance	Spec.	Sens.
Comm. Dataset (N=152)	CBR	81	26	6	39	78.9%	75.7%	86.7%
	SVM	75	32	8	37	73.7%	70%	82.2%
	PSSM	81	26	11	34	75.7%	75.7%	75.6%
	11/25 Rule	101	6	30	15	76.3%	94.4%	33.3%

Table 1. A) CBR performance in the training set, and comparison with SVM, PSSM, and the 11/25 rule. B) CBR performance in the independent dataset 1, and comparison with SVM, PSSM, and the 11/25 rule. Abbreviations: Spec=Specificity; Sens=Sensitivity; CBR= Case-Based Reasoning; PSSM=Position Specific Scoring Matrices; SVM=Support Vector Machines.

Since the similarity metric of the CBR algorithm was generated and fine-tuned based on the training set, we examined the performance of the CBR tool in other independent datasets. Additionally, we compared the sensitivity, specificity, and overall concordance in this unseen dataset with other algorithms investigated in this study. Results are shown in Table 1B. In this test, CBR outperformed all other methods in both sensitivity and overall accuracy, achieving a sensitivity of 86.7% compared to 82.2% for SVM, and 75.6% for PSSM. The sensitivity of the 11/25 rule remained very low, missing two thirds of the X4-tropic viruses.

In order to examine the robustness of the CBR tool and the case library for tropism prediction in sub-optimal conditions, a group of 128 subtype C V3 sequences was used to test the tool's predictive power. The results are shown in Table 2A. Given that the case library is comprised of predominantly subtype B samples, the CBR performed well with a specificity of 80.5%, sensitivity of 69.6% and an overall concordance of 76.6%. The CBR performance in this subtype C set was also compared to the same SVM and PSSM models, as well as the 11/25 rule. While the specificity of all these methods was very high (>95%), the sensitivity to detect X4 usage was inadequate, missing almost half of X4-tropic samples in the dataset. Moreover, to demonstrate the artificial intelligence capability of the CBR tool, and the ease of learning from new experiences, we added the subtype C sequences to the case library and evaluated the performance. The results improved substantially as shown in Table 2B, with a specificity of 84.1%, sensitivity of 73.9% and an overall concordance of 80.5%, a 6% increase.

Table 2A: Comparison of performance in subtype C dataset

	Prediction Method	TN	FP	FN	TP	Concordance	Spec.	Sens.
Subtype C (N=128)	CBR	66	16	14	32	76.6%	80.5%	69.6%
	SVM	78	4	20	26	78.9%	95.1%	56.5%

PSSM	**79**	**3**	**35**	**11**	**70.3%**	**96.3%**	**23.9%**
11/25 Rule	**81**	**1**	**34**	**12**	**72.7%**	**98.8%**	**26.1%**

Table 2B: Comparison of performance in subtype C dataset using an enhanced case library

	Prediction Method	**TN**	**FP**	**FN**	**TP**	**Concordance**	**Spec.**	**Sens.**
Subtype C (N=128)	**CBR Using Case Lib+C**	**69**	**13**	**12**	**34**	**80.5%**	**84.1%**	**73.9%**

Table 2. A) CBR performance on the subtype C dataset, and comparison with SVM, PSSM, and the 11/25 rule. B) CBR performance on the subtype C dataset when including subtype C samples into the case library. Abbreviations: Spec=Specificity; Sens=Sensitivity; CBR= Case-Based Reasoning; PSSM=Position Specific Scoring Matrices; SVM=Support Vector Machines.

Finally, we explored the feasibility and possible benefits of combining these bioinformatics methods [26]. Due to the poor performance of the 11/25 rule, only CBR, SVM, and PSSM were included in this analysis. Tropism predictions made by each method were examined, and the true positive and true negative calls were investigated in the form of a Venn diagram. Figure 3A shows the calls made within X4 using (X4 or DM tropic) viruses, and 3B displays the predictions within R5-tropic subset. As shown in Figure 3A, among 326 X4 using viruses there are 23 correctly called positive by solely CBR, compared to 1 by SVM and 3 by PSSM. In contrast, among 406 R5 viruses, 27 were correctly identified by SVM, and 2 by PSSM that were falsely called positive by CBR (Figure 3B). The non-overlapping sets of true positive and negative samples led us to believe that combining the predictions from different algorithms may improve the classification accuracy. We have implemented 2 ensemble algorithms by voting, utilizing the 3 classifiers that showed reasonable performance in this study (CBR, SVM, and PSSM) and predicting X4 usage if: (1) at least 1 method calls the sample X4 (*anyX4*), and (2) if ≥2 out of 3 predict X4 (*majorityX4*). Concordance of *anyX4* and *majorityX4* with ESTA were 77% and 87%, with sensitivity of 94% and 79% and specificity of 66% and 92%, respectively.

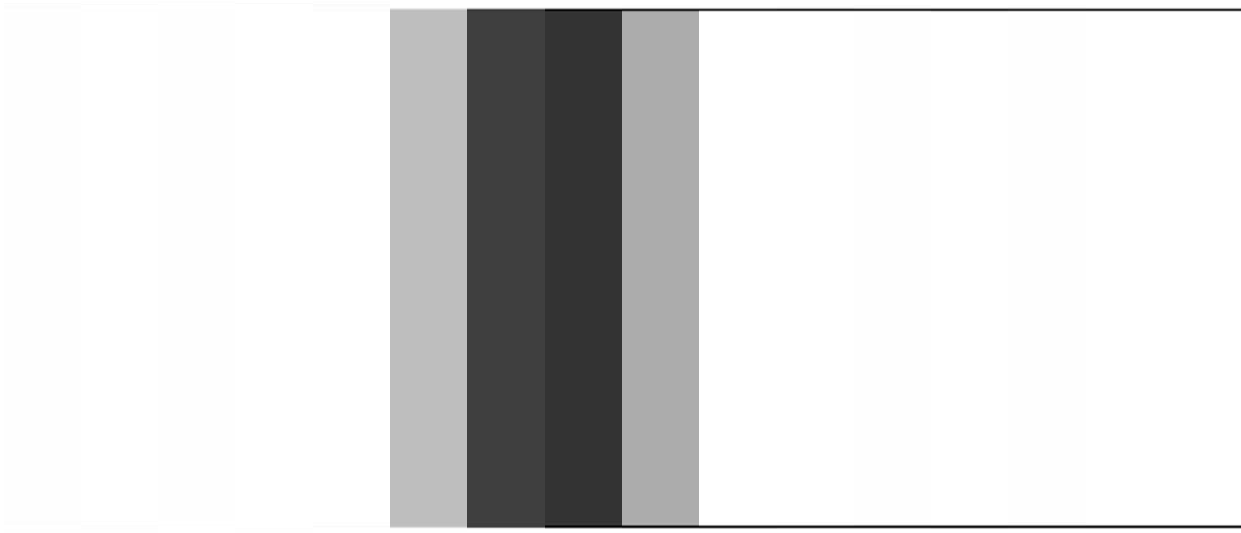

Figure 3: Venn diagram for tropism calls made by CBR, SVM, and PSSM models: **A)** within X4 using viruses (X4 or DM tropic), **B)** within R5 tropic viruses. CBR= Case-Based Reasoning; PSSM=Position Specific Scoring Matrices; SVM=Support Vector Machines.

Discussion

Performing *in-silico* prediction of HIV-1 co-receptor usage on the basis of the V3 loop is a challenging task due to the high variability of the viral envelope. We present a novel approach utilizing the case-based reasoning (CBR) technique to perform genotypic tropism prediction. Additionally, the performance of several bioinformatics techniques utilized as research tools or in the clinical practice are investigated and compared to CBR. In the training set, CBR achieved a higher sensitivity (89.6%) than SVM (85.3%), PSSM (80.7%), or the 11/25 rule (36.2%). The specificity of the CBR tool (73.2%) was lower than SVM (90.1%) and the 11/25 rule (96.5%), but comparable to PSSM (75.1%). The CBR adaptation strategy was adjusted to have high X4 sensitivity, since it would be important to identify patients who are not good candidates for CCR5 antagonist therapy and may have better treatment options. As a trade-off to higher sensitivity, CBR achieved lower specificity. When these methods were evaluated in an independent dataset, sensitivity to detect X4 usage was considerably better for CBR (86.7%), compared to SVM (82.2%), PSSM (75.6%), and the 11/25 rule (33.3%), while specificity dropped or remained at a comparable level as the training set. Since models such as SVM and PSSM are developed by generating a mathematical fit based on the training set, lower performance is expected when the model is applied on unseen datasets. CBR seems to have an advantage in that aspect since the core knowledge base is stored as a set of cases with their solution, and even though the similarity metric and adaptation strategy need to be fit to the training set, the main database is not extracted into a different format where sub-optimal extrapolation might be performed.

Diagnostic accuracy of CBR depends on the distribution of the study population stored in the case library, and can be improved by including a large spectrum of V3 sequences with diverse characteristics into the library. In our study, the accuracy of the algorithm for predicting tropism in a subtype C dataset improved by 6% when a set of samples with subtype C were added to the case library. This also demonstrates the ease of implementing and maintaining a CBR system. Existing databases of matched phenotype and genotype

can be utilized as a case library. Furthermore, the CBR algorithm can learn from new experiences by adding informative cases to the library. Instance based reasoning (IBR) which is a subclass of the case-based reasoning family has been implemented by Prosperi et al [9], but hasn't demonstrated improved performance as compared to SVM. This may be at least partly due to the implementation of the IBR system in that study utilizing Euclidean distance rather than a weighted similarity metric which was used in our CBR and allowed us to take advantage of the detailed significance levels of the features associated with the viral tropism.

We examined the granularity of the tropism predictions made by the CBR, SVM, and PSSM algorithms. Among 732 samples in the training set, we found 24 DM- tropic viruses that none of the algorithms could correctly identify as X4-using. We speculate that the X4 determination for these viruses may lie outside of the V3 loop [27, 28]. Additional studies with the entire gp160 sequence are necessary in order to confirm and identify other regions that influence co-receptor usage. For the remaining cases, correct tropism predictions were made by each individual algorithm that were false negative or positive by others. This suggests that each method has unique strength, and therefore, applying sophisticated boosting techniques [29, 30] may lead to better results. We implemented simple ensemble algorithms by voting. While using *anyX4* improved the sensitivity to detect X4-using viruses to 94%, the specificity took a hit and was reduced to 66%. The reverse happened for *majorityX4* that achieved improved specificity (91%), but decreased sensitivity (79%). It would be worth investigating other, more complex meta algorithms such as bagging [31] and decision trees, utilizing the scores generated by all these techniques and combining them to improve the accuracy of the predictions.

We demonstrated that different sites and amino-acid substitutions in the V3 loop as well as additional physiochemical and sequence attributes influence the co-receptor tropism differently. We found evidence, as examined by Fisher's Exact test, that mutations 7Y, 7K, 8I, 9K, 11R, and 30V are amongst amino-acid changes strongly associated with X4 usage (odds ratio > 20, Bonferroni corrected p-value<0.001). Additionally, we have shown that amino-acid and certain types of nucleotide mixtures occur substantially more within DM samples, which is expected given the inherent nature of Dual/Mixed viruses. We also evaluated peptide statistics extracted from the V3 sequence, and found that increased total charge, isoelectric point, and basic values, as well as decreased value measured in the small grouping are strongly associated with X4 tropism. In general, we found a set of sequence, physiochemical, and molecular characteristics of the V3 peptide that correlated with tropism. Here, we present this biologically relevant data, and were able to leverage this information and utilized the additional properties of the V3 loop to better assess similarity in the context of tropism. Some of these additional sequence characteristics could not easily be incorporated into the SVM and PSSM models, which may contribute to the lower accuracy of these models compared to CBR. Additional studies correlating the predictions with the clinical outcome of patients who had undergone CCR5 antagonist therapy would be required to assess the algorithm as a predictor of response [32].

In conclusion, case-based reasoning could be utilized as a genotypic tropism prediction algorithm. We were able to achieve improved sensitivity and specificity in independent datasets when comparing CBR with other bioinformatics predictors, in particular, SVM, PSSM, and the 11/25 rule. Further prospective studies are necessary in order to evaluate the feasibility of applying a CBR based tropism prediction tool prior to utilization in a clinical setting.

Author's information:

Haddad (MH): Bioinformatics/Biostatistics, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 616-3645
Evans (ME): Bioinformatics/Biostatistics, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 624-4181
Paquet (AP): Bioinformatics/Biostatistics, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 624-4106
Huang (WH): Research & Development, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 866-7229
Frantzell (AF): Research & Development, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 866-7449
Toma (JT): Research & Development, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 624-4282
Napolitano (LN): Clinical Research, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 866-7433
Coakley (EC): Clinical Research, Monogram Biosciences Inc., South San Francisco, CA 94080, USA
Whitcomb (JW): Operations, Monogram Biosciences Inc., South San Francisco, CA 94080, USA, Tel: (650) 866-7433
Goetz (MBG): VA Greater Los Angeles Healthcare System and David Geffen School. of Med. at UCLA

Acknowledgement: The authors would like to acknowledge the Monogram Biosciences clinical reference and R&D laboratories for performance of all phenotype and genotype assays. The CPCRA and INSIGHT networks are funded by the National Institutes of Health (U01 AI-42170, U01 AI-46362, U01 AI-68641).

Author Discloser statement: MH, ME, AP, WH, LN, AF, JT, ES, CP, JW, and EC received salary in the past 5 years from Monogram Biosciences.

Authors' Contributions: Mojgan Haddad designed and led the study, performed statistical analysis, and wrote the paper. Mark Evans performed the bioinformatics analysis, generated the parameters used for the study, and co-wrote the Methods section. Agnes Paquet performed statistical analysis and coding. Wei Huang, Arne Frantzell, and Jon Toma performed the genotypic experiments. Laura Napalitano and Eoin Coakley participated in the clinical data analysis. Eric Stawiski contributed to the bioinformatics analysis and coding. Matthew Goetz managed one of the clinical cohorts used in this study. Christos Petropoulos and Jeannette Whitcomb managed the phenotypic and genotypic experiments in the research and clinical labs.

References

1. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, *et al.* **Identification of a major co-receptor for primary isolates of HIV-1**. *Nature* 1996,381:661-666.
2. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, *et al.* **Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity**. *Antimicrob Agents Chemother* 2005,49:4721-4732.
3. Lin NH, Kuritzkes DR. **Tropism testing in the clinical management of HIV-1 infection**. *Curr Opin HIV AIDS* 2009,4:481-487.
4. Pillai S, Good B, Richman D, Corbeil J. **A new perspective on V3 phenotype prediction**. *AIDS Res Hum Retroviruses* 2003,19:145-149.
5. Sing T, Low AJ, Beerenwinkel N, Sander O, Cheung PK, Domingues FS, *et al.* **Predicting HIV coreceptor usage on the basis of genetic and clinical covariates**. *Antiviral therapy* 2007,12:1097-1106.
6. Resch W, Hoffman N, Swanstrom R. **Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks**. *Virology* 2001,288:51-62.
7. Masso M, Vaisman, II. **Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage**. *BMC Bioinformatics*,11:494.
8. Dybowski JN, Heider D, Hoffmann D. **Prediction of co-receptor usage of HIV-1 from genotype**. *PLoS computational biology* 2010,6:e1000743.
9. Prosperi MCF, Fanti I, Ulivi G, Micarelli A, De Luca A, Zazzi M. **Robust supervised and unsupervised statistical learning for HIV type 1 coreceptor usage analysis**. *AIDS research and human retroviruses* 2009,25:305-314.
10. Jensen MA, Li FS, van 't Wout AB, Nickle DC, Shriner D, He HX, *et al.* **Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences**. *J Virol* 2003,77:13376-13388.
11. Briggs DR, Tuttle DL, Sleasman JW, Goodenow MM. **Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages)**. *AIDS* 2000,14:2937-2939.
12. De Jong JJ, De Ronde A, Keulen W, Tersmette M, Goudsmit J. **Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution**. *J Virol* 1992,66:6777-6780.
13. Jensen MA, van 't Wout AB. **Predicting HIV-1 coreceptor usage with sequence analysis**. *AIDS Rev* 2003,5:104-112.
14. Low AJ, Dong W, Chan D, Sing T, Swanstrom R, Jensen M, *et al.* **Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates**. *AIDS* 2007,21:F17-24.
15. Kolodner J. **Case-Based Reasoning**. San Mateo, CA: Morgan Kaufmann Publishers, Inc.; 1993.

16. Schank RC. **Dynamic Memory: A Theory of Learning in Computers and People**. New York: Cambridge Univ. Press; 1982.
17. Kolodner JLaS, R.L. **The mediator: a case study of a case-based problem solver**: School of Information and Computer Science, Georgia Institute of Technology; 1988.
18. Haddad M, Adlassnig KP, Porenta G. **Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams**. *Artif Intell Med* 1997,9:61-78.
19. Goetz MB, Leduc R, Kostman JR, Labriola AM, Lie Y, Weidler J, *et al.* **Relationship between HIV coreceptor tropism and disease progression in persons with untreated chronic HIV infection**. *J Acquir Immune Defic Syndr* 2009,50:259-266.
20. Whitcomb JM, Huang W, Fransen S, Limoli K, Toma J, Wrin T, *et al.* **Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism**. *Antimicrob Agents Chemother* 2007,51:566-575.
21. Rice P, Longden I, Bleasby A. **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends Genet.* 2000,16:276-277.
22. Eddy SR. **Multiple alignment using hidden Markov models**. *Proc Int Conf Intell Syst Mol Biol* 1995,3:114-120.
23. Eddy SR. **Hidden Markov models**. *Curr Opin Struct Biol* 1996,6:361-365.
24. Eddy SR. **Profile hidden Markov models**. *Bioinformatics* 1998,14:755-763.
25. Vandekerckhove LP, Wensing AM, Kaiser R, Brun-Vezinet F, Clotet B, De Luca A, *et al.* **European guidelines on the clinical management of HIV-1 tropism testing**. *Lancet Infect Dis*,11:394-407.
26. Chueca N, Garrido C, Alvarez M, Poveda E, de Dios Luna J, Zahonero N, *et al.* **Improvement in the determination of HIV-1 tropism using the V3 gene sequence and a combination of bioinformatic tools**. *J Med Virol* 2009,81:763-767.
27. Huang W, Eshleman SH, Toma J, Fransen S, Stawiski E, Paxinos EE, *et al.* **Coreceptor tropism in human immunodeficiency virus type 1 subtype D: high prevalence of CXCR4 tropism and heterogeneous composition of viral populations**. *J Virol* 2007,81:7885-7893.
28. Huang W, Toma J, Fransen S, Stawiski E, Reeves JD, Whitcomb JM, *et al.* **Coreceptor tropism can be influenced by amino acid substitutions in the gp41 transmembrane subunit of human immunodeficiency virus type 1 envelope protein**. *J Virol* 2008,82:5584-5593.
29. Saigo H, Uno T, Tsuda K. **Mining complex genotypic features for predicting HIV-1 drug resistance**. *Bioinformatics* 2007,23:2455-2462.
30. Schapire R. E. SY. **Improved Boosting Algorithms Using Confidence-rated Predictions**. *Machine Learning* 1999,37:40.
31. Breiman L. **Bagging predictors**. *Machine Learning* 1996,24:123–140.
32. McGovern RA, Thielen A, Mo T, Dong W, Woods CK, Chapman D, *et al.* **Population-based V3 genotypic tropism assay: a retrospective analysis using screening samples from the A4001029 and MOTIVATE studies**. *AIDS*,24:2517-2525.