BioSystems 137 (2015) 12–19

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Identification of dual-tropic HIV-1 using evolved neural networks

Gary B. Fogel[a], Susanna L. Lamers[b,*], Enoch S. Liu[a], Marco Salemi[c], Michael S. McGrath[d]

[a] Natural Selection, Inc., San Diego, CA 92121, United States
[b] Bioinfoexperts, LLC, Thibodaux, LA 70302, United States
[c] University of Florida, Department of Pathology and Laboratory Medicine, Gainesville, FL 32610, United States
[d] University of California at San Francisco, Department of Laboratory Medicine and The AIDS and Cancer Specimen Resource, San Francisco, CA 94143, United States

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form
24 September 2015
Accepted 26 September 2015
Available online 28 September 2015

Keywords:
HIV-1 coreceptor
Viral tropism
Artificial neural network
Evolutionary computation
HIV phenotype

ABSTRACT

Blocking the binding of the envelope HIV-1 protein to immune cells is a popular concept for development of anti-HIV therapeutics. R5 HIV-1 binds CCR5, X4 HIV-1 binds CXCR4, and dual-tropic HIV-1 can bind either coreceptor for cellular entry. R5 viruses are associated with early infection and over time can evolve to X4 viruses that are associated with immune failure. Dual-tropic HIV-1 is less studied; however, it represents functional antigenic intermediates during the transition of R5 to X4 viruses. Viral tropism is linked partly to the HIV-1 envelope V3 domain, where the amino acid sequence helps dictate the receptor a particular virus will target; however, using V3 sequence information to identify dual-tropic HIV-1 isolates has remained difficult. Our goal in this study was to elucidate features of dual-tropic HIV-1 isolates that assist in the biological understanding of dual-tropism and develop an approach for their detection. Over 1559 HIV-1 subtype B sequences with known tropisms were analyzed. Each sequence was represented by 73 structural, biochemical and regional features. These features were provided to an evolved neural network classifier and evaluated using balanced and unbalanced data sets. The study resolved R5X4 viruses from R5 with an accuracy of 81.8% and from X4 with an accuracy of 78.8%. The approach also identified a set of V3 features (hydrophobicity, structural and polarity) that are associated with tropism transitions. The ability to distinguish R5X4 isolates will improve computational tropism decisions for R5 vs. X4 and assist in HIV-1 research and drug development efforts.

1. Introduction

Approximately 34 million people are infected with HIV-1 worldwide and more than 1.1 million live with HIV-1 infection in the United States (Hall et al., 2008). While combined antiretroviral therapy (cART) has increased the lifespan of HIV-1-infected individuals, cART does not clear viral infection (Alexaki et al., 2008; Aquaro et al., 1998). Its success, however, has supported the current mission to cure HIV-1 disease (Stevenson, 2014). However, patients can acquire resistance to cART, which results from specific genetic mutations in the viral genome (Barrie et al., 1996; Gulnik et al., 1995). Resistance creates the need for patients to change medication with drugs that attack the virus in different ways (Cortez and Maldarelli, 2011). HIV-1 variation, whether occurring naturally or in direct response to cART, also influences escape from immune surveillance (Coffin, 1995), disease pathogenesis (Salemi et al., 2005, 2009), development of viral reservoirs (Salemi et al., 2009), and a wide spectrum of diseases associated with metabolic disorder (Fitch et al., 2013; Bernstein et al., 2006; Estrada and Portilla, 2011), neurological disorder (Anthony et al., 2005) and cancer (Salemi et al., 2009; Ng and McGrath, 1998; Lamers et al., 2010).

In order to enter immune cells, HIV-1 first binds a CD4 cellular receptor and then another receptor, usually CCR5 (R5) or CXCR4 (X4) (MMW). These co-receptors are present on the surface of both macrophages and T-cells (Goodenow and Collman, 2006); however, the complex preference of HIV-1 for specific co-receptors varies under different conditions, such as passage in cell cultures (Goodenow and Collman, 2006) or the state of the immune system (Gorry et al., 2005). Still, as co-receptor binding is required for successful HIV-1 infection, exploiting the process for therapeutic intervention remains a popular concept. Entry inhibitors (EIs) are a class of drugs designed specifically to block coreceptors, thereby limiting the ability of HIV-1 to infect new cells. Currently, Maraviroc (MacArthur and Novak, 2008; Dorr et al., 2005) blocks CCR5 while drug development continues for drugs that target CXCR4. Administering these drugs requires matching the therapy to the appropriate receptor in use by HIV-1. Furthermore, HIV-1 can evolve to acquire

* Corresponding author.
E-mail address: susanna@bioinfox.com (S.L. Lamers).

http://dx.doi.org/10.1016/j.biosystems.2015.09.007
0303-2647/

resistance to EIs (Starr-Spires and Collman, 2002) in a manner similar to that of other antiretroviral drugs. Therefore, the ability to monitor which receptor(s) are being used for viral entry is critical for appropriate treatment options based on the current and anticipated evolution of the virus.

R5 HIV-1 plays a crucial role in the transmission and establishment of HIV-1 (Gorry and Ancuta, 2011). Classically, the emergence of X4 viruses has been associated with the progression of acquired immune deficiency syndrome (AIDS, 1990; Moyle et al., 2005). However, many patients never evolve X4 HIV-1 and instead evolve a highly macrophage-tropic HIV-1 with enhanced tropism for R5 (Gorry et al., 2005) or "dual-tropic" HIV-1 (R5X4) (Tasca et al., 2008), which can use both co-receptors for cellular entry (Goodenow and Collman, 2006; Gorry and Ancuta, 2011; Loftin et al., 2011; Robertson et al., 2000). Dual tropic HIV-1 is an interesting intermediate in that it is less efficient in binding either receptor than R5 or X4, but still allows for viral entry (Tasca et al., 2008). Furthermore, recent views suggest that X4 emergence is not associated with a highly pathogenic virus, but rather is the result of reduced host immune efficiency (Tasca et al., 2008), which permits the accumulation of viral diversity (Mild et al., 2013), resulting in a wider spectrum of co-receptor usage. In this light, the ability to identify R5 and R5X4 viruses prior to X4 emergence with high accuracy would accelerate the study of HIV-1 evolution (Campbell et al., 2014), the staging of disease progression (Gorry et al., 2005; Berger et al., 1999; Broder and Collman, 1997; Murakami and Yamamoto, 2000; Weber et al., 2006; Clapham and McKnight, 2002; Moore et al., 2004; Weiss, 2002; Doms et al., 2000), and the development of appropriate personalized therapies for those infected (Starr-Spires and Collman, 2002; Katzenstein, 2003).

One of the principal determinants of HIV-1 interactions with R5 or X4 cellular coreceptors is the V3 domain of the HIV-1 envelope protein (Resch et al., 2001). Co-receptor selection of viral isolates in this region is influenced by amino acid substitutions, insertions, and deletions (Hoffman et al., 2002). HIV-1 V3 sequence data sets with known phenotypes combined with artificial neural networks (ANNs) or other machine learning strategies, primarily with a focus on backpropagation for training, have been applied to predict R5 vs. X4 co-receptor usage with reasonable success (Resch et al., 2001; Ioannidis et al., 2003; Wang and Larder, 2003; Brumme et al., 2004; Milich et al., 1993). However, the continued inability to identify dual-tropic R5X4, could account for decreased sensitivity and specificity of these methods if true R5X4 sequences are actually being misclassified as either R5 or X4 for model development and testing. Further, in the case of artificial neural networks (ANNs), although backpropagation is a common strategy for ANN optimization, convergence is only guaranteed to a locally optimal solution. A different approach to ANN optimization makes use of evolutionary computation to discover weight assignments and/or evolve the ANN architecture itself. Evolved neural networks (ENNs) (Fogel, 2008; Fogel et al., 1990; Porto et al., 1995; Yao, 1999; Kohl and Miikkulainen, 2011) have been applied with success to a wide variety of biochemical data mining problems (Fogel, 2008; Hecht et al., 2008; Hecht and Fogel, 2007; Lamers et al., 2008) and afford the researcher with the opportunity not only optimize a model for input-output mapping but to examine which feature combinations taken from a larger set of possible features are most relevant for high accuracy classification.

In a previous publication, we presented the first use of ENNs to classify HIV-I co-receptor use (Lamers et al., 2008). That initial research was based on a small public set of 149 HIV-1 V3 loop sequences (77 R5, 31 R5X4, and 41 X4 sequences) from a variety of HIV-1 subtypes with known tropisms. 9 biochemical features for each of 35 amino acid positions and 2 additional V3-domain-level features were calculated. Fully connected feed-forward ENNs were used to map the features for each sequence to co-receptor usage classification using increasingly larger feature sets as inputs. The effort not only produced useful classifiers but also helped identify feature combinations that were important for classification. ENNs were trained to classify R5 sequences from X4 sequences, and additional ENNs were trained to classify R5X4 sequences from either R5 or X4 sequences. This approach led to a mean classification accuracy of 88.9% for R5 vs. X4 and a mean classification accuracy of 75.5% for R5X4 vs. R5 or X4. This initial approach demonstrated strong potential for correctly classifying dual-tropic HIV-1 using an expanded set of sequences and features. In this paper, we used a larger database of over 1559 sequences and a broader assortment of 73 features to derive classifiers for four separate tropism decisions (R5 vs. X4, R5 vs. R5X4, X4 vs. R5X4, and R5 vs. R5X4 vs. X4). While many strategies for nonlinear machine learning could be applied to this problem such as support vector machines (SVMs), we specifically chose to use ENNs for this work in order to compare results to our previous effort as described above. Further, we evaluated the effects of balanced vs. unbalanced data sets on model performance. The results provide an indication of important features associated with tropism classification as well as improved detection of dual-tropic viruses.

2. Methods

V3 Loop Sequences. Publically available V3 loop sequences (relative to HXB2 positions 7110-7217) for HIV-1 subtype B were downloaded from the HIV Database at the Los Alamos National Laboratory (http://www.hiv.lanl.gov/content/index) and translated into amino acid sequences. The search criteria limited each data set to either "only CCR5," "only CXCR4," or "only R5X4.". The resulting database consisted of 3452 R5 sequences, 197 X4 sequences, and 545 R5X4 sequences. The sequences were aligned using ClustalW within the MEGA5 sequence analysis package (Tamura et al., 2011) and then manually edited to correct for any obvious alignment errors. Identical sequences were removed. The final alignment used for feature generation below contained 1223 unique R5 sequences, 241 unique R5X4 sequences, and 95 unique X4 sequences. No limitation was put on the phenotype or culture method in order to preserve sequence diversity.

2.1. Feature generation and processing

While many studies have focused on characteristics at specific sites in the V3 loop relative to tropism (Briggs et al., 2000; Wilkin and Gulick, 2012; Delgado et al., 2012), we took a different unbiased approach using both site-specific and regional characteristics. For each sequence, 73 features were calculated for each of the 40 alignment positions (Supplemental Table 1). These features were selected using the available tropism literature and also through resources such as ExPASy programs ProtScale and ProtParam (www.expasy.org) (Wilkins et al., 1999). While some features were position-dependent (e.g., glycosylation at specific positions), the remaining features were calculated for all positions and, to determine if regional features were associated with tropism, the alignment was further reduced to smaller segments: the 5′ end of the alignment (positions 9–14), the 3′ end of the alignment (positions 22–28) and lastly for a second region at the 3′ end of the alignment (positions 31–37). This process resulted in ~3.000 possible feature-positions that could be provided as input to a model for classification. Linear regression was then used to determine which features were useful in separating the tropism classes independently for each of four decisions (R5 vs. R5X4, R5 vs. X4, R5X4 vs. X4, and R5 vs. R5X4 vs. X4) and to reduce the number of features to those with highest correlation to tropism class decision for each decision type.